TRANSACTIONS
OF THE
AMERICAN INSTITUTE
OF
ELECTRICAL ENGINEERS

VOLUME 80
1961

PART I. COMMUNICATION AND ELECTRONICS

PUBLISHED BY THE
AMERICAN INSTITUTE OF ELECTRICAL ENGINEERS
345 EAST 47TH STREET
NEW YORK 17, N. Y., U. S. A.

Copyright 1962 by the American Institute of Electrical Engineers
Printed in U. S. A. by The Mack Printing Company, Easton, Pa.

Preface

The AIEE *Transactions* for 1961 (volume 80) is published in three parts: Part I. Communication and Electronics; Part II. Applications and Industry; and Part III. Power Apparatus and Systems. The papers in each of the three parts are classified according to subject matter as follows:

Part I. Communication and Electronics

Aero-Space Instrumentation
Basic Sciences
Broadcasting
Communication Switching Systems
Communication Theory
Computing Devices
Data Communication
Education
Electrical Insulation
Electrical Techniques in Medicine and Biology
Electron Tubes
Electronic and High-Frequency Instruments
Electronic Circuits and Systems
Electronic Transformers
Electronics
Fundamental Electrical Standards
Indicating and Integrating Instruments
Magnetic Amplifiers
Nuclear Instrumentation
Nucleonics
Radio Communication Systems
"Radio Engineering" Review
"Radio Engineering and Electronic Physics" Review
Recording and Controlling Instrumentation
Semiconductor Rectifiers
Solid State Devices
Space Communications
Special Instruments and Auxiliary Apparatus
"Telecommunications" Review
Telegraph Systems
Telemetering
Wire Communication Systems

Part II. Applications and Industry

Aero-Space Transportation
Cement Industry
Chemical Industry
Domestic and Commercial Applications
Electric Heating
Electric Welding
Feedback Control Systems
General Industry Applications
Industrial and Commercial Power Systems
Industrial Control
Industrial Power Rectifiers
Land Transportation
Machine Tools Industry
Marine Transportation
Metal Industry
Mining Industry
Petroleum Industry
Production and Application of Light
Rubber and Plastic Industries
Textile Industry

Part III. Power Apparatus and Systems

Insulated Conductors
Power Generation
Power System Communications
Protective Devices
Relays
Rotating Machinery
Substations
Switchgear
System Engineering
Transformers
Transmission and Distribution

New Technical Activities
Research
Safety
Standards

Each part has been indexed separately in the back of that particular part. The three parts are not cross-referenced; hence the user should determine first whether the subject matter of the paper desired is in the field of communication and electronics, applications and industry, or power apparatus and systems. Papers are listed in the subject index under several key words in the titles. The original numbers assigned to the papers are given in the author index. Volume 80 contains the technical papers and related discussions presented at these meetings:

1. Winter General Meeting, New York, N. Y., January 29–February 3, 1961.
2. South East–South Central District Meeting, New Orleans, La., April 5–7, 1961.
3. Northeastern District Meeting, Hartford, Conn., May 17–19, 1961.
4. Summer General Meeting, Ithaca, N. Y., June 18–23, 1961.
5. Joint Automatic Control Conference, Boulder, Colo., June 28–30, 1961.
6. Pacific General Meeting, Salt Lake City, Utah, August 23–25, 1961.
7. Fall General Meeting, Detroit, Mich., October 15–20, 1961.

Statements and opinions given in the papers and discussions published in *Transactions* are the expressions of the contributors for which the American Institute of Electrical Engineers assumes no responsibility.

Communications and Electronics Index for 1961

1. Subject Index

AM, FM, And TV Transmitting Station Parameters, Automatic Logging of. Ehrenberg..........199–204
Adder, A 4-Megacycle 24-Bit Checked Binary. Homan..........443–50
Adders and Other Logic, Analog and Threshold Building Blocks for Variable-Radix. Sklansky..........289–95
Air in the Electric Wind of the Corona Discharge, Movement of. Robinson..........143–50
Algebraic Equations, Automatic Plotting of Characteristic Curves and Analog Solution of. Mitrovic..........701–03; disc. 703
Algebras in Switching Theory, E-. Semon..........265–9
Algorithm for Determining Minimal Normal Forms of an Incomplete Truth Function, An. Mott....73–6
Algorithms for Logical Design. Ewing, Roth, Wagner..........450–8
Ampere Mechanical Switch, A 20,000-. Newgard..........96–8
Ampere-Turn Gain of Ordinary Saturable Reactors, An Experimental Study on the. Geyger.....173–8
Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity, Transistor. Mulligan, Shamis..........335–40
Amplifier with a Saturable Reactor Controlling On Time, A New Magnetic-Controlled Rectifier Power. Morgan..........152–5
Amplitude Distortion in Transistor Feedback Amplifiers. Mulligan..........326–35
Analog and Threshold Building Blocks for Variable-Radix Adders and Other Logic. Sklansky..289–95
Analog Computer Stabilization of High-Speed Magnetic Amplifiers. Greene, Saibel..........17–22
Analog of Transport Phenomena in a Thermoelectric Generator, Electric. Talaat..........427–31
Analog Simulation of Digital Computer Programs. Saucedo, Sze..........703–09
Analog-to-Digital Converters, Tentative Recommendation for the Definition and Testing of the Dynamic Performance of. Gaines, Fischer. 387–93; disc. 393
Analysis of Conventional Single-Ended Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of Transfer Functions. Nahi..........394–400
Analysis of Errors in the Calibration of Electric Instruments. Hermach..........90–5; disc. 95
Analysis of Nonlinear Systems, General Recurrence Relations in the. Wolf..........383–7
Antenna Array, The Theory of a Single-Ring Circular. Hickman, Neff, Tillman..........110–15
Antenna Matching Unit for High-Frequency Vehicular Whip. Gruber, Seward..........99–104
Approximate Representation of a Curve with Linear Segments, A Note on the. Shiva..........461–3
Artery Disease, Instrumentation for the Diagnosis of Coronary-. Moyer, Talbott..........717–21
Automatic Dialing and Electronic Selection on Railroad Message and Dispatcher Circuits. Jorgensen..........204–08
Automatic Logging of AM, FM, and TV Transmitting Station Parameters. Ehrenberg..........199–204
Automatic Plotting of Characteristic Curves and Analog Solution of Algebraic Equations. Mitrovic..........701–03; disc. 703
Automatic Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System. Warchol, Sachs, Lorentson..........77–80

B

(Basic Sciences) A Method of Calculating Gain and Phase Margins from Pole-Zero Diagrams. Schlereth..........139–43
(Basic Sciences) A Note on the Approximate Representation of a Curve with Linear Segments. Shiva..........461–3
(Basic Sciences) Automatic Plotting of Characteristic Curves and Analog Solution of Algebraic Equations. Mitrovic..........701–03; disc. 703
(Basic Sciences) Design of High Coercivity Permanent Magnets, Exposed to External Fields. Schindler..........423–7
(Basic Sciences) Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops. Sutton, Hurwitz, Poritsky.....687–95
(Basic Sciences) General Recurrence Relations in the Analysis of Nonlinear Systems. Wolf..........383–7
(Basic Sciences) Magnetization and Pull Characteristics of Mating Magnetic Reeds. Peek......372–80
(Basic Sciences) Maximization and Minimization of Complicated Multivariable Functions. Bromberg..........725–30; disc. 730
(Basic Sciences) The Determination of the Hysteresis Curve for Thick Tape Cores. Della Torre, Dentella..........359–62
(Basic Sciences) The Synthesis of General Parameter Insertion Loss Filters Using a Digital Computer. Temes..........181–6
Behavior of a Thermoelectric Generator, The Initial Transient. Stremler, Gray..........367–72
Bell System 1A Echo Suppressors, A Tone Disabler for. Bugbee..........596–600
Billing and Collection System Using Teletypewriter Techniques, A New. Henches..........560–6
Binary Adder, A 4-Megacycle 24-Bit Checked. Homan..........443–50
Binary Controls for Error Control. Peterson..........648–52; disc. 652
Binary Sieve Method, Boolean Prime Implicants by the. Hall..........709–13
Blocks for Logical Design, Forcing Circuitry: Sequential Building. Meade..........522–31
Boolean Prime Implications by the Binary Sieve Method. Hall..........709–13
Bridge-Lifting Problem and Some Methods of Solution, The Subscriber Loop. Hochgraf......240–5
British JERC Development of Electronic Switching, The. Flowers, Forshaw..........208–17
(Broadcasting) Automatic Logging of AM, FM, and TV Transmitting Station Parameters. Ehrenberg..........199–204
(Broadcasting) Noise and Intermodulation Problems in Multichannel Closed-Circuit Television Systems. Collins, Williams..........486–91
Buffer Store for Data Transmission, A. Baker, Froehlich..........566–70
Building Blocks for Logical Design, Forcing Circuitry: Sequential. Meade..........522–31
Building Blocks for Variable-Radix Adders and Other Logic, Analog and Threshold. Sklansky..........289–95
Business Community, Centrex Service: A New Design for Customer Group Telephone Service in the Modern. Shea..........474–82

C

Cable Facilities, Synthesis of Networks Simulating Impedance of Loaded. DeMonte..........623–7
Cable Irregularities on the Operation of Telephone Repeaters, Effects of. McCaron..........217–14
Cables, Electrical Uniformity and Deviations Control in Polyethylene-Insulated Multipaired. Roberts..........570–6; disc. 576
Calculating Gain and Phase Margins from Pole-Zero Diagrams, A Method of. Schlereth..........139–43
Calibration of Electric Instruments, An Analysis of Errors in the. Hermach..........90–5; disc. 95
Careers, Educating Electrical Engineers for Professional. Linke..........551–3; disc. 553
Carrier System—Applications, The 81-A Exchange Trunk. Cool, Walker..........617–23
Carrier System—Repeaters and Regulation, The 81-A Exchange Trunk. Gilmore, Griffith..........606–10
Carrier System—Signaling, The 81-A Exchange Trunk. Tuck, Melvin..........610–17
Carrier System, The 81-A Exchange Trunk. Ferguson, Harp..........601–06
Carrier System, Type T-12S-Tr Transistorized Short-Haul. Ishihara, Yamamoto, Fujita, Murakami..........193–9
Centrex Service: A New Design for Customer Group Telephone Service in the Modern Business Community. Shea..........474–82
Challenge, Motivation Through. Johnson, Clement..........554–6
Channel, Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst. Cowell, Burton..........577–85
Channel Due to End Current Loops, Electrical and Pressure Losses in a Magnetohydrodynamic. Sutton, Hurwitz, Poritsky..........687–95
Channel Multiplex System, A Completely Transistorized 600-. Sparks..........121–6
Character Recognition System, Rotating Raster. Weeks..........353–9
Characteristic Curves and Analog Solution of Algebraic Equations, Automatic Plotting of, Mitrovic..........701–03; disc. 703
Characteristics of Mating Magnetic Reeds, Magnetization and Pull. Peek..........372–80
Charging of Nonspherical Particles in a Corona Discharge, The. Smith, Penney..........340–6
Chemical Processing Model Considered as a Nonlinear System, Considerations in the Design of a Dynamic Control System for the Generalized. Beecher, Gould..........413–23
Cineradiography, X-Ray Tube with Grid for Control of Exposures in. Rogers..........713–16
Circuit and Equipment Impedances as They Affect Transmission Performance, Telephone. Radue..........126–34
Circuit for Pulse-Height to Pulse-Width Conversion, A Low-Level Linear Rundown. Endsley, Grannemann, Summers..........150–2
Circuitry: Sequential Building Blocks for Logical Design, Forcing. Meade..........522–31
Circuits, Automatic Dialing and Electronic Selection on Railroad Message and Dispatcher. Jorgensen..........204–08
Circuits, Commutation and Destructive Oscillation in Diode. Somos..........162–72; disc. 172
Circuits, Design of Diode-Switch High-Sensitivity Trigger. Vojinovic..........458–61
Circuits, Minimization of Multiple-Output Switching. Polansky..........67–73
Circuits of Electrostatic Precipitators, Determination and Simulation of the Equivalent. Thomas, Drenning, Williams..........315–19; disc. 319
Circular Antenna Array, The Theory of a Single-Ring. Hickman, Neff, Tillman..........110–15
Closed-Circuit Television Systems, Noise and Intermodulation Problems in Multichannel. Collins, Williams..........486–91
Code for Quaternary Data Transmission, An Error-Correction. Scherer..........231–6
Code Converter, Transistor Morse to Teleprinter. Cuniff, Theall..........178–81
Coded Feedback Communication Systems. Metzner, Morgan..........643–7
Codes with Retransmission on a Gilbert Burst Channel, Computer Simulation of the Use of Group. Cowell, Burton..........577–85
Coercivity Permanent Magnets, Exposed to External Fields, Design of High. Schindler..........423–7
Collection System Using Teletypewriter Techniques, A New Billing and. Henches..........560–6
(Committee Report) Report on Behavior of Magnetic Materials Under Extreme Environmental Conditions. (T-121)..........58–60
(Communication Switching Systems) A Low-Cost Number Identifier for Small Telephone Systems. Redington..........117–21
(Communication Switching Systems) Centrex Service: A New Design for Customer Group Telephone Service in the Modern Business Community. Shea..........474–82
(Communication Switching Systems) Features of an Electronic Crosspoint PABX. Van Bosse, DeCicco..........471–4
(Communication Switching Systems) Logical Control of an Electronic Crosspoint PABX. Sanders..........496–501
(Communication Switching Systems) The British Development of Electronic Switching. Flowers, Forshaw..........208–17
(Communication Switching Systems) Transmission Network of an Electronic Crosspoint PABX. Kowalik..........491–6
Communication Systems, Coded Feedback. Metzner, Morgan..........643–7
(Communication Theory) An Error-Correction Code for Quaternary Data Transmission. Scherer..........231–6
(Communication Theory) Binary Controls for Error Control. Peterson..........648–52; disc. 652
(Communication Theory) Modulation and Signal Selection for Digital Data Systems. Lerner..........661–9
Communications Electronics, A Stitched-Wiring Process for Miniaturized. Coffin, Johnson..........682–6; disc. 686
Commutation, A Silicon-Controlled Rectifier Inverter with Improved. McMurray, Shattuck.....531–42
Commutation and Destructive Oscillation in Diode Circuits. Somos..........162–72; disc. 172
Comparison of Computers, A. Curl...542–6; disc. 546

Comparison of Control Computers, Results of Simulation. Sendzuk..........547–50; disc. 550
Completely Transistorized 600-Channel Multiplex System, A. Sparks..........121–6
Computer Program for Preparing Wiring Diagrams. Kirby, Rosenthal..........509–13
Computer Programs, Analog Simulation of Digital. Saucedo, Sze..........703–09
Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel. Cowell, Burton..........577–85
Computer Stabilization of High-Speed Magnetic Amplifiers, Analog. Greene, Saibel..........17–22
Computers, A Comparison of. Curl...542–6; disc. 546
Computers, Results of Simulation Comparison of Control. Sendzuk..........547–50; disc. 550
(Computing Devices) A 4-Megacycle 24-Bit Checked Binary Adder. Homan..........443–50
(Computing Devices) Algorithms for Logical Design. Ewing, Roth, Wagner..........450–8
(Computing Devices) An Algorithm for Determining Minimal Normal Forms of an Incomplete Truth Function. Mott..........73–6
(Computing Devices) Analog and Threshold Building Blocks for Variable-Radix Adders and Other Logic. Sklansky..........289–95
(Computing Devices) Boolean Prime Implicants by the Binary Sieve Method. Hall..........709–13
Computing Devices) Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System. Beecher, Gould..........413–23
(Computing Devices) E-Algebras in Switching Theory. Semon..........265–9
(Computing Devices) Forcing Circuitry— Sequential Building Blocks for Logical Design. Meade..........522–31
(Computing Devices) Large-Scale On-Line Data Processing Systems. Levine..........505–09
(Computing Devices) Logical Design by Regression-Matrix Logic VII. Schubert..........380–3
(Computing Devices) Minimization of Multiple-Output Switching Circuits. Polansky..........67–73
Computing Devices) Rotating Raster Character Recognition System. Weeks..........353–9
Computing Devices) Tentative Recommendation for the Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters. Gaines, Fischer..........387–93; disc. 393
Conditions, Report on Behavior of Magnetic Materials Under Extreme Environmental. (Committee Report)..........58–60
Conference Rooms, A Speakerphone System for Large. Huggler..........237–40; disc. 240
Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System. Beecher, Gould..........413–23
Contaminants: The Inclined-Plane Liquid-Contaminant Test, Surface Electrical Failure in the Presence of. Mathes, McGowan..281–6; disc. 286
Control Computers, Results of Simulation Comparison of. Sendzuk..........547–50; disc. 550
Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C Networks. Kadri..........43–8
Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System, Considerations in the Design of a Dynamic. Beecher, Gould..........413–23
Control Tungsten Lamp Loads, Problems Encountered in Applying the Silicon-Controlled Rectifier to. Daugherty..........400–02
Controllable Low-Ripple Constant-Current Source, A. Bonnema, Slemon..........3–8
Conversion, A Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width. Endsley, Grannemann, Summers..........150–2
Converter Circuits by Means of R-C Networks, Control of Frequency and Phase Displacement in Transistor. Kadri..........43–8
Converter Power Supply, A Magnetically Regulated D-C to D-C. Sager..........513–18
Converter, Transistor Morse to Teleprinter Code. Cunniff, Theall..........178–81
Converter, Tunnel Diode D-C Power. Storm, Shattuck..........347–52
Converters, Tentative Recommendation for the Definition and Testing of the Dynamic Performance of Analog-to-Digital. Gaines, Fischer..........387–93; disc. 393
Core Modulation Integrator, A Saturable-. Barker, Gruodis..........13–17
Cores, An Improved Model for Flux Reversal in Ni-Fe. Leliakov, Friedlaender..........23–6
Cores, Detailed Measurements of Slow Magnetization Processes in Tape-Wound. Brownell, Barker..........402–12
Cores, Initiation of Flux Reversal in Magnetic-Amplifier. Friedlaender, Leliakov..........269–72
Cores, On the Process of Flux Reversal in Multiaperature Ferrite. Rowe, Slemon..........431–8
Cores, The Determination of the Hysteresis Curve for Thick Tape. Della Torre, Dentella..........359–52
Corona Discharge, Movement of Air in the Electric Wind of the. Robinson..........143–50
Corona Discharge, The Charging of Nonspherical Particles in a. Smith, Penney..........340–6
Coronary-Artery Disease, Instrumentation for the Diagnosis of. Moyer, Talbott..........717–21
Crosspoint PABX, Features of an Electronic. Van Bosse, DeCicco..........417–4
Crosspoint PABX, Logical Control of an Electronic. Sanders..........496–501
Crosspoint PABX, Transmission Network of an Electronic. Kowalik..........491–6
Current Source, A Controllable Low-Ripple Constant-. Bonnema, Slemon..........3–8
Currents, A Magnetic Device for High-Speed Sensing of Small. Baldwin..........1–3
Curve with Linear Segments, A Note on the Approximate Representation of a. Shiva..........461–3
Curves and Analog Solution of Algebraic Equations, Automatic Plotting of Characteristic. Mitrovic..........701–03; disc. 703
Customer Group Telephone Service in the Modern Business Community, Centrex Service: A New Design for. Shea..........474–82

D

(Data Communication) Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel. Cowell, Burton....577–85
(Data Communication) Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines. Kelly, Mercurio, Willard..........585–90
Data Extraction, Magnetics in Doppler Signal. Metz, Fay..........33–43
Data Gathering and Transmission System for Project EHV. Abetti, Caverly, Davis, Whepley....272–81
Data Processing Systems, Large-Scale On-Line. Levine..........505–09
Data System for Use Over the Switched Telephone Network, A 1,650-Bit-Per-Second. Brand, Carter..........652–61
Data System Tests Using Simulated Transmission Impairments. Andrews..........590–6
Data Systems, Modulation and Signal Selection for Digital. Lerner..........661–9
Data Transmission, A Buffer Store for. Baker, Froehlich..........566–70
Data Transmission, An Error-Correction Code for Quaternary. Scherer..........231–6
Data Transmission System for Telephone Circuits, The Design of an Error-Free. Reiffen, Schmidt, Yudkin..........224–31
Data Transmission Through Noisy Media—A Systems Approach, Reliable. Melas..........501–04
D-C Digital Voltmeters, A General Description of. Stansbury..........465–70
D-C Power Converter, Tunnel Diode. Storm, Shattuck..........347–52
D-C to D-C Converter Power Supply, A Magnetically Regulated. Sager..........513–18
Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters, Tentative Recommendation for the. Gaines, Fischer..........387–93; disc. 393
Description of D-C Digital Voltmeters, A General. Stansbury..........465–70
Design, Algorithms for Logical. Ewing, Roth, Wagner..........450–8
Design by Regression—Matrix Logic VII, Logical. Schubert..........380–3
Design of a 2-Way Radio Channel for Maintenance of Telephone Company Services. MacDonald..........639–42
Design of an Error-Free Data Transmission System for Telephone Circuits, The. Reiflen, Schmidt, Yudkin..........224–31
Design of Diode-Switch High-Sensitivity Trigger Circuits. Vojinovic..........458–61
Design of High Coercivity Permanent Magnets, Exposed to External Fields. Schindler..........423–7
Design of Resistive Temperature Compensation by Single and Multiple Thermistor Networks. Farhi, Groves..........246–53
Design of the Generator Voltage Regulator, The. Van Emden..........438–43
Destructive Oscillation in Diode Circuits, Commutation and. Somos..........162–72; disc. 172
Detailed Measurements of Slow Magnetization Processes in Tape-Wound Cores. Brownell, Barker...402–12
Determination and Simulation of the Equivalent Circuits of Electrostatic Precipitators. Thomas, Drenning, Williams..........315–19; disc. 319
Determination of the Hysteresis Curve for Thick Tape Cores, The. Della Torre, Dentella..........359–62
Device Structure, A Fatigue-Free Silicon. Green..........186–91; disc. 191
Devices, On the Elimination of Null in Modulating and Demodulating. Jones..........135–9; disc. 139
Diagnosis of Coronary-Artery Disease, Instrumentation for the. Moyer, Talbott..........717–21
Diagrams, A Method of Calculating Gain and Phase Margins from Pole-Zero. Schlereth..........139–43
Diagrams, Computer Program for Preparing Wiring. Kirby, Rosenthal..........509–13
Dialing and Electronic Selection on Railroad Message and Dispatcher Circuits, Automatic. Jorgensen..........204–08
Digital Computer Programs, Analog Simulation of. Saucedo, Sze..........703–09
Digital Computer, The Synthesis of General Parameter Insertion Loss Filters Using a. Temes..........181–6
Digital Data Systems, Modulation and Signal Selection for. Lerner..........661–9
Digital Magnetic Recording, Optical Read-out of. Miyata..........53–8
Digital Telemetering Techniques. Brothman, Brothman, Reiser..........81–90
Digital Trophospheric Scatter. Poppe, Vadasz..669–74
Digital Voltmeters, A General Description of D-C. Stansbury..........465–70
Diode Circuits, Commutation and Destructive Oscillation in. Somos..........162–72; disc. 172
Diode D-C Power Converter, Tunnel. Storm, Shattuck..........347–52
Diode-Switch High-Sensitivity Trigger Circuits, Design of. Vojinovic..........458–61
Direct-Voltage Ripple, Effective Value of. Diebold..........61–6; disc. 66
Disabler for Bell System 1A Echo Suppressors, A Tone. Bugbee..........596–600
Discharge, Movement of Air in the Electric Wind of the Corona. Robinson..........143–50
Discharge, The Charging of Nonspherical Particles in a Corona. Smith, Penney..........340–6
Discharges Preceding Sparkover at Low-Voltage Gradients, Pulse. Penney, Craig..........156–62
Dispatcher Circuits, Automatic Dialing and Electronic Selection on Railroad Message and. Jorgensen..........204–08
Distortion and Distortion Measuring, Telegraph. Wusteney..........104–09
Distortion in Transistor Feedback Amplifiers, Amplitude. Mulligan..........326–35
Doppler Signal Data Extraction, Magnetics in. Metz, Fay..........33–43
Dosage, Technical Solutions to the Problem of Reducing Patient. Jacobs..........721–5
Dropouts on Telephone Lines, Two Measuring Techniques for the Investigation of Impulse Noise and. Kelly, Mercurio, Willard..........585–90
Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System, Considerations in the Design of a. Beecher, Gould..........413–23

E

E-Algebras in Switching Theory. Semon..........265–9
Echo Suppressors, A Tone Disabler for Bell System 1A. Bugbee..........596–600
Educating Electrical Engineers for Professional Careers. Linke..........551–3; disc. 553
(Education) Motivation Through Challenge. Johnson, Clement..........554–6
Effective Value of Direct-Voltage Ripple. Diebold..........61–6; disc. 66
Effects of Cable Irregularities on the Operation of Telephone Repeaters. McCaron..........217–14
EHV, Data Gathering and Transmission System for Project. Abetti, Caverly, Davis, Whepley..........272–81
81-A Exchange Trunk Carrier System, The. Ferguson, Harp..........601–06
81-A Exchange Trunk Carrier System—Applications, The. Cool, Walker..........617–23
81-A Exchange Trunk Carrier System—Repeaters and Regulation, The. Gilmore, Griffith..........606–10
81-A Exchange Trunk Carrier System—Signaling, The. Tuck, Melvin..........610–17
Electric Analog of Transport Phenomena in a Thermoelectric Generator. Talaat..........427–31
Electric Instruments, An Analysis of Errors in the Calibration of. Hermach..........90–5; disc. 95
Electric Wind of the Corona Discharge, Movement of Air in the. Robinson..........143–50
Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops. Sutton, Hurwitz, Poritsky..........687–95
Electrical Engineers for Professional Careers, Educating. Linke..........551–3; disc. 553
Electrical Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test, Surface. Mathes, McGowan....281–6; disc. 286
(Electrical Techniques in Medicine & Biology) Instrumentation for the Diagnosis of Coronary-Artery Disease. Moyer, Talbott..........717–21

(Electrical Techniques in Medicine & Biology) Technical Solutions to the Problem of Reducing Patient Dosage. Jacobs....721–5
(Electrical Techniques in Medicine & Biology) X-Ray Tube with Grid for Control of Exposures in Cineradiography. (61-34)....713–16
Electrical Uniformity and Deviations Control in Polyethylene-Insulated Multipaired Cables. Roberts....570–6; disc. 576
Electromechanical Pulser, Linear Analysis of the. Pande....309–15
Electronic Crosspoint PABX, Features of an. Van Bosse, DeCicco....471–4
Electronic Crosspoint PABX, Logical Control of an. Sanders....496–501
Electronic Crosspoint PABX, Transmission Network of an. Kowalik....491–6
Electronic Method for High-Speed Teleprinters, A Proposed. Biswas....115–17
Electronic Selection on Railroad Message and Dispatcher Circuits, Automatic Dialing and. Jorgensen....204–08
Electronic Switching, The British JERC Development of. Flowers, Forshaw....208–17
Electronic Transformers, Ultrahigh-Temperature:
  Part I. Rippen, Harms Walters..295–302; disc. 302
  Part II. Walters....302–09
(Electronics) A Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion. Endsley, Grannemann, Summers....150–2
Electronics, A Stitched-Wiring Process for Miniaturized Communications. Coffin, Johnson....682–6; disc. 686
(Electronics) Amplitude Distortion in Transistor Feedback Amplifiers. Mulligan....326–35
(Electronics) Design of Diode-Switch High-Sensitivity Trigger Circuits. Vojinovic....458–61
(Electronics) Movement of Air in the Electric Wind of the Corona Discharge. (61–90)....143–50
(Electronics) On the Elimination of Null in Modulating and Demodulating Devices. Jones....135–9; disc. 139
(Electronics) Pulse Discharges Preceding Sparkover at Low Voltage Gradients. Penney, Craig....156–62
(Electronics) The Charging of Nonspherical Particles in a Corona Discharge. Smith, Penney....340–6
(Electronics) The Design of the Generator Voltage Regulator. Van Emden....438–43
(Electronics) Transistor Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity. Mulligan, Shamis....335–40
(Electronics) Transistor Morse to Teleprinter Code Convert. Cunniff, Theall....178–81
Electrostatic Precipitators, Determination and Simulation of the Equivalent Circuits of. Thomas, Drenning, Williams....315–19; disc. 319
Elimination of Null in Modulating and Demodulating Devices, On the. Jones....135–9; disc. 139
Environmental Conditions, Report on Behavior of Magnetic Materials Under Extreme. (Committee Report)....58–60
Equations, Automatic Plotting of Characteristic Curves and Analog Solution of Algebraic. Mitrovic....701–03; disc. 703
Equipment Impedances as They Affect Transmission Performance, Telephone Circuit and. Radue....126–34
Equivalent Circuits of Electrostatic Precipitators, Determination and Simulation of the. Thomas, Drenning, Williams....315–19; disc. 319
Errata....464
Error Control, Binary Controls for. Peterson....648–52; disc. 652
Error-Correction Code for Quaternary Data Transmission, An. Scherer....231–6
Error-Free Data Transmission System for Telephone Circuits, The Design of an. Reiffen, Schmidt, Yudkin....224–31
Errors in the Calibration of Electric Instruments, An Analysis of. Hermach....90–5; disc. 95
Exchange Trunk Carrier System—Applications, The 81-A. Cool, Walker....617–23
Exchange Trunk Carrier System—Repeaters and Regulation, The 81-A. Gilmore, Griffith....606–10
Exchange Trunk Carrier System—Signaling, The 81-A. Tuck, Melvin....610–17
Exchange Trunk Carrier System, The 81-A. Ferguson, Harp....601–06
Experimental Study on the Ampere-Turn Gain of Ordinary Saturable Reactors, An. Geyger..173–8
Exposures in Cineradiography, X-Ray Tube with Grid for Control of. Rogers....713–16

F

Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test, Surface Electrical. Mathes, McGowan....281–6; disc. 286
Fatigue-Free Silicon Device Structure, A. Green....186–91; disc. 191
Features of an Electronic Crosspoint PABX. Van Bosse, DeCicco....471–4
Feedback Amplifiers, Amplitude Distortion in Transistor. Mulligan....326–35
Feedback Communication Systems, Coded. Metzner, Morgan....643–7
Ferrite Cores, On the Process of Flux Reversal in Multiaperature. Rowe, Slemon....431–8
Ferrites, Neutron and Gamma Irradiation of Sqme Square-Loop and Microwave. Moss, Kooi, Baldwin....362–7
Filters Using a Digital Computer, The Synthesis of General Parameter Insertion Loss. Temes...181–6
Flow Graph Theory for Linear Time-Variable Systems, Signal. Stear, Stubberud....695–701
Flux Reversal in Magnetic-Amplifier Cores, Initiation of. Friedlaender, Leliakov....269–72
Flux Reversal in Multiaperature Ferrite Cores, On the Process of. Rowe, Slemon....431–8
Flux Reversal in Ni-Fe Cores, An Improved Model for. Leliakov, Friedlaender....23–6
Forcing Circuitry: Sequential Building Blocks for Logical Design. Meade....522–31
Forms of an Incomplete Truth Function, An Algorithm for Determining Minimal Normal. Mott....73–6
4-Megacycle 24-Bit Checked Binary Adder, A. Homan....443–50
Frequencies of Even Subharmonics, Relaxation Oscillations at Multiple. Prywes....48–52
Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C Networks, Control of. Kadri....43–8
Functions, Maximization and Minimization of Complicated Multivariable. Bromberg....725–30; disc. 730

G

Gain and Phase Margins from Pole-Zero Diagrams, A Method of Calculating. Schlereth....139–43
Gain and Static and Dynamic Sensitivity, Transistor Amplifier Stages with Prescribed. Mulligan, Shamis....335–40
Gain of Ordinary Saturable Reactors, An Experimental Study on the Ampere-Turn. Geyger....173–8
Gamma Irradiation of Some Square-Loop and Microwave Ferrites, Neutron and. Moss, Kooi, Baldwin....362–7
General Description of D-C Digital Voltmeters, A. Stansbury....465–70
General Recurrence Relations in the Analysis of Nonlinear Systems. Wolf....383–7
Generator, Electric Analog of Transport Phenomena in a Thermoelectric. Talaat....427–31
Generator, The Initial Transient Behavior of a Thermoelectric. Stremler, Gray....367–72
Generator Voltage Regulator, The Design of the. Van Emden....438–43
Gilbert Burst Channel, Computer Simulation of the Use of Group Codes with Retransmission on a. Cowell, Burton....577–85
Gradients, Pulse Discharges Preceding Sparkover at Low Voltage. Penney, Craig....156–62
Graph Theory for Linear Time-Variable Systems, Signal Flow. Stear, Stubberud....695–701
Group Codes with Retransmission on a Gilbert Burst Channel, Computer Simulation of the Use of. Cowell, Burton....577–85
Group Telephone Service in the Modern Business Community, Centrex Service: A New Design for Customer. Shea....474–82

H

High Coercivity Permanent Magnets, Exposed to External Fields, Design of. Schindler....423–7
High-Frequency Vehicular Whip, Antenna Matching Unit for. Gruber, Seward....99–104
High-Speed Magnetic Amplifiers, Analog Computer Stabilization of. Greene, Saibel....17–22
High-Speed Sensing of Small Currents, A Magnetic Device for. Baldwin....1–3
High-Speed Teleprinters, A Proposed Electronic Method for. Biswas....115–17
Hysteresis Curve for Thick Tape Cores, The Determination of the. Della Torre, Dentella....359–62

I

Identifier for Small Telephone Systems, A Low-Cost Number. Redington....117–21
Impairments, Data System Tests Using Simulated Transmission. Andrews....590–6
Impedance of Ink Recorders, Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective. Geyger....8–13
Impedance of Loaded Cable Facilities, Synthesis of Networks Simulating. DeMonte....623–7
Impedances as They Affect Transmission Performance, Telephone Circuit and Equipment. Radue.126–34
Improved Model for Flux Reversal in Ni-Fe Cores, An. Leliakov, Friedlaender....23–6
Impulse Noise and Dropouts on Telephone Lines, Two Measuring Techniques for the Investigation of. Kelly, Mercurio, Willard....585–90
Incoherent Scattering, A New Method of Space Exploration by Radar, Using. Laaspere....557–60
(Indicating and Integrating Instruments) A General Description of D-C Digital Voltmeters. Stansbury....465–70
Inductive-Resistive Load Including the Derivation of Transfer Functions, Analysis of Conventional Single-Ended Magnetic Amplifiers with. Nahi....394–400
Initial Transient Behavior of a Thermoelectric Generator, The. Stremler, Gray....367–72
Initiation of Flux Reversal in Magnetic-Amplifier Cores. Friedlaender, Leliakov....269–72
Ink Recorders, Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of. Geyger....8–13
Insertion Loss Filters Using a Digital Computer, The Synthesis of General Parameter. Temes....181–6
Instrument for Time Saving and Better Measurement in the Speech Band, The Level Tracer, an. Niedereder, Sanders....674–82
Instrumentation for the Diagnosis of Coronary-Artery Disease. Moyer, Talbott....717–21
Instruments, An Analysis of Errors in the Calibration of Electric. Hermach....90–5; disc. 95
Instruments, Taut Band Suspension Switchboard. Nyes, Thomander, MacIndoe, Roerty....258–64
Insulated Multipaired Cables, Electrical Uniformity and Deviations Control in Polyethylene-. Roberts....570–6; disc. 576
Integrator, A Saturable-Core Modulation. Barker, Gruodis....13–17
Intermodulation Problems in Multichannel Closed-Circuit Television Systems, Noise and. Collins, Williams....486–91
Inverter with Improved Commutation, A Silicon-Controlled Rectifier. McMurray, Shattuck....531–42
Inverters, Self-Locking Polyphase Transistor-Magnetic. Campling, Bennett....26–33
Irradiation of Some Square-Loop and Microwave Ferrites, Neutron and Gamma. Moss, Kooi, Baldwin....362–7

J

JERC Development of Electronic Switching, The British. Flowers, Forshaw....208–17

L

Lamp Loads, Problems Encountered in Applying the Silicon-Controlled Rectifier to Control Tungsten. Daugherty....400–02
Large-Scale On-Line Data Processing Systems. Levine....505–09
Level Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band, The. Niedereder, Sanders....674–82
Linear Analysis of the Electromechanical Pulser. Pande....309–15
Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion, A Low-Level. Endsley, Grannemann, Summers....150–2
Linear Segments, A Note on the Approximate Representation of a Curve with. Shiva....461–3
Linear Time-Variable Systems, Signal Flow Graph Theory for. Stear, Stubberud....695–701
Load Including the Derivation of Transfer Functions, Analysis of Conventional Single-Ended Magnetic Amplifiers with Inductive-Resistive. Nahi....394–400
Locking Polyphase Transistor-Magnetic Inverters, Self-. Campling, Bennett....26–33
Logging of AM, FM, and TV Transmitting Station Parameters, Automatic. Ehrenberg....199–204
Logical Control of an Electronic Crosspoint PABX. Sanders....496–501
Logical Design, Algorithms for. Ewing, Roth, Wagner....450–8
Logical Design by Regression—Matrix Logic VII. Schubert....380–3
Logical Design, Forcing Circuitry: Sequential Building Blocks for. Meade....522–31
Loop Bridge-Lifting Problem and Some Methods of Solution, The Subscriber. Hochgraf....240–5
Loss Filters Using a Digital Computer, The Synthesis of General Parameter Insertion. Temes....181–6
Losses in a Magnetohydrodynamic Channel Due to End Current Loops, Electrical and Pressure. Sutton, Hurwitz, Poritsky....687–95
Low-Cost Number Identifier for Small Telephone Systems, A. Redington....117–21
Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion, A. Endsley, Grannemann, Summers....150–2

Low-Ripple Constant-Current Source, A Controllable. Bonnema, Slemon....3–8
Low Voltage Gradients, Pulse Discharges Preceding Sparkover at. Penney, Craig....156–62

M

Magnetic-Amplifier Cores, Initiation of Flux Reversal in. Friedlaender, Leliakov....269–72
(Magnetic Amplifiers) A Controllable Low-Ripple Constant-Current Source. Bonnema, Slemon....3–8
(Magnetic Amplifiers) A Saturable-Core Modulation Integrator. Barker, Gruodis....13–17
(Magnetic Amplifiers) An Experimental Study on the Ampere-Turn Gain of Ordinary Saturable Reactors. Geyger....773–8
(Magnetic Amplifiers) An Improved Model for Flux Reversal in Ni-Fe Cores. Leliakov, Friedlaender....23–6
Magnetic Amplifiers, Analog Computer Stabilization of High-Speed. Greene, Saibel....17–22
(Magnetic Amplifiers) Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C Networks. Kadri....43–8
Magnetic Amplifiers for Changing the Effective Impedance of Ink Recorders, Single-Ended Self-Balancing. Geyger....8–13
(Magnetic Amplifiers) On the Process of Flux Reversal in Multiaperature Ferrite Cores. Rowe, Slemon....431–8
(Magnetic Amplifiers) Relaxation Oscillations at Multiple Frequencies of Even Subharmonics. Prywes....48–52
(Magnetic Amplifiers) Tunnel Diode D-C Power Converter. Storm, Shattuck....347–52
Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of Transfer Functions, Analysis of Conventional Single-Ended. Nahi....394–400
Magnetic-Controlled Rectifier Power Amplifier with a Saturable Reactor Controlling On Time, A New. Morgan....152–5
Magnetic Device for High-Speed Sensing of Small Currents, A. Baldwin....1–3
Magnetic Inverters, Self-Locking Polyphase Transistor-. Campling, Bennett....26–33
Magnetic Materials Under Extreme Environmental Conditions, Report on Behavior of. (Committee Report)....58–60
Magnetic Recording, Optical Readout of Digital. Miyata....53–8
Magnetically Regulated D-C to D-C Converter Power Supply, A. Sager....513–18
Magnetics in Doppler Signal Data Extraction. Metz, Fay....33–43
Magnetization and Pull Characteristics of Mating Magnetic Reeds. Peek....372–80
Magnetization Processes in Tape-Wound Cores, Detailed Measurements of Slow. Brownell, Barker....402–12
Magnetohydrodynamic Channel Due to End Current Loops, Electrical and Pressure Losses in a. Sutton, Hurwitz, Poritsky....687–95
Magnetometer, A Pulse-Position Modulator-Type. Stricker, Wulkan....253–8
Magnets, Exposed to External Fields, Design of High Coercivity Permanent. Schindler....423–7
Maintenance of Telephone Company Services, Design of a 2-Way Radio Channel for. MacDonald....639–42
Margins from Pole-Zero Diagrams, A Method of Calculating Gain and Phase. Schlereth....139–43
Mating Magnetic Reeds, Magnetization and Pull Characteristics of. Peek....372–80
Matrix Logic VII, Logical Design by Regression—Schubert....380–3
Maximization and Minimization of Complicated Multivariable Functions. Bromberg..725–30; disc. 730
Measurements of Slow Magnetization Processes in Tape-Wound Cores, Detailed. Brownell, Barker....402–12
Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines, Two. Kelly, Mercurio, Willard....585–90
Measuring, Telegraph Distortion and Distortion. Wusteney....104–09
Mechanical Switch, A 20,000-Ampere. Newgard..96–8
Media—A Systems Approach, Reliable Data Transmission Through Noisy. Melas....501–04
Megacycle 24-Bit Checked Binary Adder, A 4-. Homan....443–50
Method of Calculating Gain and Phase Margins from Pole-Zero Diagrams, A. Schlereth....139–43
Method of Space Exploration by Radar, Using Incoherent Scattering, A New. Laaspere....557–60
Microwave Ferrites, Neutron and Gamma Irradiation of Some Square-Loop and. Moss, Kooi, Baldwin....362–7
Microwave System, Santa Barbara County 6,000-Mc. Lentz....627–34
Miniaturized Communications Electronics, A Stitched-Wiring Process for. Coffin, Johnson....682–6; disc. 686
Minimal Normal Forms of an Incomplete Truth Function, An Algorithm for Determining. Mott...73–6
Minimization of Multiple-Output Switching Circuits. Polansky....67–73
Model for Flux Reversal in Ni-Fe Cores, An Improved. Leliakov, Friedlaender....23–6
Modulating and Demodulating Devices, On the Elimination of Null in. Jones....135–9; disc. 139
Modulation and Signal Selection for Digital Data Systems. Lerner....661–9
Modulation Integrator, A Saturable-Core. Barker, Gruodis....13–17
Modulator-Type Magnetometer, A Pulse-Position. Stricker, Wulkan....253–8
Morse to Teleprinter Code Converter, Transistor. Cunniff, Theall....178–81
Motivation Through Challenge. Johnson, Clement....554–6
Movement of Air in the Electric Wind of the Corona Discharge. Robinson....143–50
Multiaperature Ferrite Cores, On the Process of Flux Reversal in. Rowe, Slemon....431–8
Multichannel Closed-Circuit Television Systems, Noise Intermodulation Problems in. Collins, Williams....486–91
Multipaired Cables, Electrical Uniformity and Deviations Control in Polyethylene-Insulated. Roberts....570–6; disc. 576
Multiple-Output Switching Circuits, Minimization of. Polansky....67–73
Multiple Thermistor Networks, Design of Resistive Temperature Compensation by Single and. Farhi, Groves....246–53
Multiplex System, A Completely Transistorized 600-Channel. Sparks....121–6
Multivariable Functions, Maximization and Minimization of Complicated. Bromberg....725–30; disc. 730

N

Network, A 1,650-Bit-Per-Second Data System for Use Over the Switched Telephone. Brand, Carter....652–61
Network of an Electronic Crosspoint PABX, Transmission. Kowalik....491–6
Networks, Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C. Kadri....43–8
Networks, Design of Resistive Temperature Compensation by Single and Multiple Thermistor. Farhi, Groves....246–53
Networks Simulating Impedance of Loaded Cable Facilities, Synthesis of. DeMonte....623–7
Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites. Moss, Kooi, Baldwin....362–7
(New York City) Centrex Service: A New Design for Customer Group Telephone Service in the Modern Business Community. Shea....474–82
Ni-Fe Cores, An Improved Model for Flux Reversal in. Leliakov, Friedlaender....23–6
Noise and Dropouts on Telephone Lines, Two Measuring Techniques for the Investigation of Impulse. Kelly, Mercurio, Willard....585–90
Noise and Intermodulation Problems in Multichannel Closed-Circuit Television Systems. Collins, Williams....486–91
Noisy Media—A Systems Approach, Reliable Data Transmission Through. Melas....501–04
Nonlinear System, Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a. Beecher, Gould....413–23
Nonlinear Systems, General Recurrence Relations in the Analysis of. Wolf....383–7
Note on the Approximate Representation of a Curve with Linear Segments, A. Shiva....461–3
(Nucleonics) Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites. Moss, Kooi, Baldwin....362–7
Null in Modulating and Demodulating Devices, On the Elimination of. Jones....135–9; disc. 139
Number Identifier for Small Telephone Systems, A Low-Cost. Redington....117–21

O

On-Line Data Processing Systems, Large-Scale. Levine....505–09
1,650-Bit-Per-Second Data System for Use Over the Switched Telephone Network, A. Brand, Carter....652–61
Optical Read-out of Digital Magnetic Recording. Miyata....53–8
Oscillation in Diode Circuits, Commutation and Destructive. Somos....162–72; disc. 172
Oscillations at Multiple Frequencies of Even Subharmonics, Relaxation. Prywes....48–52
Oscillator with Current Transformer, A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor. Jackson....518–22
Outage Location by Means of Very-High-Frequency Radio Systems, Remote. Cooper....635–9
Output Transistor Oscillator with Current Transformer, A 30-Volt 10-Ampere Regulated Power Supply Using the Variable. Jackson....518–22

P

PABX, Features of an Electronic Crosspoint. Van Bosse, DeCicco....471–4
PABX, Logical Control of an Electronic Crosspoint. Sanders....496–501
PABX, Transmission Network of an Electronic Crosspoint. Kowalik....491–6
Parameter Insertion Loss Filters Using a Digital Computer, The Synthesis of General. Temes....181–6
Parameters, Automatic Logging of AM, FM, and TV Transmitting Station. Ehrenberg....199–204
Particles in a Corona Discharge, The Charging of Nonspherical. Smith, Penney....340–6
Patient Dosage, Technical Solutions to the Problem of Reducing. Jacobs....721–5
Performance of Analog-to-Digital Converters, Tentative Recommendation or the Definition and Testing of the Dynamic. Gaines, Fischer....387–93; disc. 393
Permanent Magnets, Exposed to External Fields, Design of High Coercivity. Schindler....423–7
Phase Displacement in Transistor Converter Circuits by Means of R-C Networks, Control of Frequency and. Kadri....43–8
Phase Margins from Pole-Zero Diagrams, A Method of Calculating Gain and. Schlereth....139–43
Phenomena in a Thermoelectric Generator, Electric Analog of Transport. Talaat....427–31
Plotting of Characteristic Curves and Analog Solution of Algebraic Equations, Automatic. Mitrovic....701–03; disc. 703
Pole-Zero Diagrams, A Method of Calculating Gain and Phase Margins from. Schlereth....139–43
Pole-Zero Techniques Applied to Voice-Frequency Telephone Lines. Fleming....482–6
Polyethylene-Insulated Multipaired Cables, Electrical Uniformity and Deviations Control in. Roberts....570–6; disc. 576
Polyphase Transistor-Magnetic Inverters, Self-Locking. Campling, Bennett....26–33
Power Amplifier with a Saturable Reactor Controlling On Time, A New Magnetic-Controlled Rectifier. Morgan....152–5
Power-Controlled Rectifiers, Switching Characteristics of Silicon. Somos:
  Part I....320–6
Power Converter, Tunnel Diode D-C. Storm, Shattuck....347–52
Power Supply, A Magnetically Regulated D-C to D-C Converter. Sager....513–18
Power Supply Using the Variable Output Transistor Oscillator with Current Transformer, A 30-Volt 10-Ampere Regulated. Jackson....518–22
Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System, Automatic. Warchol, Sachs, Lorentson....77–80
Precipitators, Determination and Simulation of the Equivalent Circuits of Electrostatic. Thomas, Drenning, Williams....315–19; disc. 319
Prescribed Gain and Static and Dynamic Sensitivity, Transistor Amplifier Stages with. Mulligan, Shamis....335–40
Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops, Electrical and. Sutton, Hurwitz, Poritsky....687–95
Prime Implicants by the Binary Sieve Method, Boolean. Hall....709–13
Problem and Some Methods of Solution, The Subscriber Loop Bridge-Lifting. Hochgraf....240–5
Problems Encountered in Applying the Silicon-Controlled Rectifier to Control Tungsten Lamp Loads. Daugherty....400–02
Problems in Multichannel Closed-Circuit Television Systems, Noise and Intermodulation. Collins, Williams....486–91
Process of Flux Reversal in Multiaperture Ferrite Cores, On the. Rowe, Slemon....431–8
Processes in Tape-Wound Cores, Detailed Measurements of Slow Magnetization. Brownell, Barker....402–12
Processing Model Considered as a Nonlinear System, Considerations in the Design of a Dynamic Control System for the Generalized Chemical. Beecher, Gould....413–23
Processing Systems, Large-Scale On-Line Data. Levine....505–09
Professional Careers, Educating Electrical Engineers for. Linke....551–3; disc. 553
Program for Preparing Wiring Diagrams, Computer. Kirby, Rosenthal....509–13

Proposed Electronic Method for High-Speed Teleprinters, A. Biswas........115–17
Pull Characteristics of Mating Magnetic Reeds, Magnetization and. Peek........372–80
Pulse Discharges Preceding Sparkover at Low Voltage Gradients. Penney, Craig........156–62
Pulse-Height to Pulse-Width Conversion, A Low-Level Linear Rundown Circuit for. Endsley, Grannemann, Summers........150–2
Pulse-Position Modulator-Type Magnetometer, A. Stricker, Wulkan........253–8
Pulser, Linear Analysis of the Electromechanical. Pande........309–15

Q

Quaternary Data Transmission, An Error-Correction Code for. Scherer........231–6

R

R-C Networks, Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of. Kadri........43–8
Radar, Using Incoherent Scattering, A New Method of Space Exploration by. Laaspere........557–60
Radio Channel for Maintenance of Telephone Company Services, Design of a 2-Way. MacDonald........639–42
(Radio Communication Systems) A New Method of Space Exploration by Radar, Using Incoherent Scattering. Laaspere........557–60
(Radio Communication Systems) Antenna Matching Unit for High-Frequency Vehicular Whip. Gruber, Seward........99–104
(Radio Communication Systems). Digital Tropospheric Scatter. Poppe, Vadasz........669–74
(Radio Communication Systems) Santa Barbara County 6,000-Mc Microwave System. Lentz........627–34
(Radio Communication Systems) The Theory of a Single-Ring Circular Antenna Array. Hickman, Neff, Tillman........110–15
~~Radio System, Automatic Precipitation~~ and Temperature Reporting Over a Very-High-Frequency. Warchol, Sachs, Lorentson........77–80
Radio Systems, Remote Outage Location by Means of Very-High-Frequency. Cooper........635–9
Railroad Message and Dispatcher Circuits, Automatic Dialing and Electronic Selection on. Jorgensen........204–08
Raster Character Recognition System, Rotating. Weeks........353–9
Read-out of Digital Magnetic Recording, Optical. Miyata........53–8
Recognition System, Rotating Raster Character. Weeks........353–9
Recorders, Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of Ink. Geyger........8–13
Recording, Optical Read-out of Digital Magnetic. Miyata........53–8
Rectifier Inverter with Improved Commutation, A Silicon-Controlled. McMurray, Shattuck........531–42
Rectifier Power Amplifier with a Saturable Reactor Controlling On Time, A New Magnetic-Controlled. Morgan........152–5
Rectifier to Control Tungsten Lamp Loads, Problems Encountered in Applying the Silicon-Controlled. Daugherty........400–02
Rectifiers, Switching Characteristics of Silicon Power-Controlled. Somos:
  Part I........320–6
Recurrence Relations in the Analysis of Nonlinear Systems, General. Wolf........383–7
Reeds, Magnetization and Pull Characteristics of Mating Magnetic. Peek........372–80
Regression—Matrix Logic VII, Logical Design by. Schubert........380–3
Regulated D-C to D-C Converter Power Supply, A Magnetically. Sager........513–18
Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer, A 30-Volt 10-Ampere. Jackson........518–22
Relaxation Oscillations at Multiple Frequencies of Even Subharmonics. Prywes........48–52
Reliable Data Transmission Through Noisy Media—A Systems Approach. Melas........501–04
Remote Outage Location by Means of Very-High-Frequency Radio Systems. Cooper........635–9
Repeaters, Effects of Cable Irregularities on the Operation of Telephone. McCaron........217–14
Report on Behavior of Magnetic Materials Under Extreme Environmental Conditions. (Committee Report)........58–60
Representation of a Curve with Linear Segments, A Note on the Approximate. Shiva........461–3
Resistive Temperature Compensation by Single and Multiple Thermistor Networks, Design of. Farhi, Groves........246–53
Results of Simulation Comparison of Control Computers. Sendzuk........547–50; disc. 550
Ring Circular Antenna Array, The Theory of a Single-. Hickman, Neff, Tillman........110–15
Ripple, Effective Value of Direct-Voltage. Diebold........61–6; disc. 66
Rotating Raster Character Recognition System. Weeks........353–9
Rundown Circuit for Pulse-Height to Pulse-Width Conversion, A Low-Level Linear. Endsley, Grannemann, Summers........150–2

S

Santa Barbara County 6,000-Mc Microwave System. Lentz........627–34
Saturable-Core Modulation Integrator, A. Barker, Gruodis........13–17
Saturable Reactor Controlling On Time, A New Magnetic-Controlled Rectifier Power Amplifier with a. Morgan........152–5
Saturable Reactors, An Experimental Study on the Ampere-Turn Gain of Ordinary. Geyger........173–8
Scatter, Digital Tropospheric. Poppe, Vadasz........669–74
Scattering, A New Method of Space Exploration by Radar, Using Incoherent. Laaspere........557–60
Segments, A Note on the Approximate Representation of a Curve with Linear. Shiva........461–3
Selection for Digital Data Systems, Modulation and Signal. Lerner........661–9
Selection on Railroad Message and Dispatcher Circuits, Automatic Dialing and Electronic. Jorgensen........204–08
Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of Ink Recorders, Single-Ended. Geyger........8–13
Self-Locking Polyphase Transistor-Magnetic Inverters. Campling, Bennett........26–33
(Semiconductor Rectifiers) A Fatigue-Free Silicon Device Structure. Green........186–91; disc. 191
(Semiconductor Rectifiers) A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer. Jackson........518–22
(Semiconductor Rectifiers) Commutation and Destructive Oscillation in Diode Circuits. Somos........162–72; disc. 172
(Semiconductor Rectifiers) Effective Value of Direct-Voltage Ripple. Diebold........61–6; disc. 66
Sensing of Small Currents, A Magnetic Device for High-Speed. Baldwin........1–3
Sequential Building Blocks for Logical Design, Forcing Circuitry: Meade........522–31
Short-Haul Carrier System, Type T-12S-Tr Transistorized. Ishihara, Yamamoto, Fujita, Murakami........193–9
Signal Data Extraction, Magnetics in Doppler. Metz, Fay........33–43
Signal Flow Graph Theory for Linear Time-Variable Systems. Stear, Stubberud........695–701
Signal Selection for Digital Data Systems, Modulation and. Lerner........661–9
Silicon-Controlled Rectifier Inverter with Improved Commutation, A. McMurray, Shattuck........531–42
Silicon-Controlled Rectifier to Control Tungsten Lamp Loads, Problems Encountered in Applying the. Daugherty........400–02
Silicon Device Structure, A Fatigue-Free. Green........186–91; disc. 191
Silicon Power-Controlled Rectifiers, Switching Characteristics of. Somos:
  Part I........320–6
Simulated Transmission Impairments, Data System Tests Using. Andrews........590–6
Simulating Impedance of Loaded Cable Facilities, Synthesis of Networks. DeMonte........623–7
Simulation Comparison of Control Computers, Results of. Sendzuk........547–50; disc. 550
Simulation of Digital Computer Programs, Analog. Saucedo, Sze........703–09
Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel, Computer. Cowell, Burton........577–85
Single and Multiple Thermistor Networks, Design of Resistive Temperature Compensation by. Farhi, Groves........246–53
Single-Ended Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of Transfer Functions, Analysis of Conventional. Nahi........394–400
Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of Ink Recorders. Geyger........8–13
Single-Ring Circular Antenna Array, The Theory of a. Hickman, Neff, Tillman........110–15
600-Channel Multiplex System, A Completely Transistorized. Sparks........121–6
6,000-Mc Microwave System, Santa Barbara County. Lentz........627–34
(Solid State Devices) Electric Analog of Transport Phenomena in a Thermoelectric Generator. Talaat........427–31
(Solid State Devices) The Initial Transient Behavior of a Thermoelectric Generator. Stremler, Gray........367–72
Solutions to the Problem of Reducing Patient Dosage, Technical. Jacobs........721–5
Space Exploration by Radar, Using Incoherent Scattering, A New Method of. Laaspere........557–60
Sparkover at Low Voltage Gradients, Pulse Discharges Preceding. Penney, Craig........156–62
Speakerphone System for Large Conference Rooms, A. Huggler........237–40; disc. 240
Speech Band, The Level Tracer, an Instrument for Time Saving and Better Measurement in the. Niedereder, Sanders........674–82
Square-Loop and Microwave Ferrites, Neutron and Gamma Irradiation of Some. Moss, Kooi, Baldwin........362–7
Stabilization of High-Speed Magnetic Amplifiers, Analog Computer. Greene, Saibel........17–22
Static and Dynamic Sensitivity, Transistor Amplifier Stages with Prescribed Gain and. Mulligan, Shamis........335–40
Stitched-Wiring Process for Miniaturized Communications Electronics. Coffin, Johnson........682–6; disc. 686
Study on the Ampere-Turn Gain of Ordinary Saturable Reactors, An Experimental. Geyger........173–8
Subharmonics, Relaxation Oscillations at Multiple Frequencies of Even. Prywes........48–52
Subscriber Loop Bridge-Lifting Problem and Some Methods of Solution, The. Hochgraf........240–5
Suppressors, A Tone Disabler for Bell System 1A Echo. Bugbee........596–600
Surface Electrical Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test. Mathes, McGowan........281–6; disc. 286
Suspension Switchboard Instruments, Taut Band. Nycz, Thomander, MacIndoe, Roerty........258–64
Switch, A 20,000-Ampere Mechanical. Newgard........96–8
Switch High-Sensitivity Trigger Circuits, Design of of Diode-. Vojinovic........458–61
Switchboard Instruments, Taut Band Suspension. Nycz, Thomander, MacIndoe, Roerty........258–64
Switched Telephone Network, A 1,650-Bit-Per-Second Data System for Use Over the. Brand, Carter........652–61
Switching Characteristics of Silicon Power-Controlled Rectifiers. Somos:
  Part I........320–6
Switching Circuits, Minimization of Multiple-Output. Polansky........67–73
Switching, The British JERC Development of Electronic. Flowers, Forshaw........208–17
Switching Theory, E-Algebras in. Semon........265–9
Synthesis of General Parameter Insertion Loss Filters Using a Digital Computer, The. Temes........181–6
Synthesis of Networks Simulating Impedance of Loaded Cable Facilities. DeMonte........623–7

T

Tape Cores, The Determination of the Hysteresis Curve for Thick. Della Torre, Dentella........359–62
Tape-Wound Cores, Detailed Measurements of Slow Magnetization Processes in. Brownell, Barker........402–12
Taut Band Suspension Switchboard Instruments. Nycz, Thomander, MacIndoe, Roerty........258–64
Technical Solutions to the Problem of Reducing Patient Dosage. Jacobs........721–5
Techniques Applied to Voice-Frequency Telephone Lines, Pole-Zero. Fleming........482–6
Techniques, Digital Telemetering. Brothman, Brothman, Reiser........81–90
Telegraph Distortion and Distortion Measuring. Wusteney........104–09
(Telegraph Systems) A New Billing and Collection System Using Teletypewriter Techniques. Henches........560–6
(Telegraph Systems) A Proposed Electronic Method for High-Speed Teleprinters. Biswas........115–17
Telemetering Techniques, Digital. Brothman, Brothman, Reiser........81–90
Telephone Circuit and Equipment Impedances as They Affect Transmission Performance. Radue........126–34
Telephone Circuits, The Design of an Error-Free Data Transmission System for. Reiffen, Schmidt, Yudkin........224–31
Telephone Company Services, Design of a 2-Way Radio Channel for Maintenance of. MacDonald........639–42
Telephone Lines, Pole-Zero Techniques Applied to Voice-Frequency. Fleming........482–6
Telephone Lines, Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on. Kelly, Mercurio, Willard........585–90
Telephone Network, A 1,650-Bit-Per-Second Data System for Use Over the Switched. Brand, Carter........652–61

Telephone Repeaters, Effects of Cable Irregularities on the Operation of. McCaron...........217–14
Telephone Service in the Modern Business Community, Centrex Service: A New Design for Customer Group. Shea...........................474–82
Telephone Systems, A Low-Cost Number Identifier for Small. Redington.......................117–21
Teleprinter Code Converter, Transistor Morse to. Cunniff, Theall...........................178–81
Teleprinters, A Proposed Electronic Method for High-Speed. Biswas...........................115–17
Teletypewriter Techniques, A New Billing and Collection System Using. Henches...............560–6
Television Systems, Noise and Intermodulation Problems in Multichannel Closed-Circuit. Collins, Williams.................................486–91
Temperature Compensation by Single and Multiple Thermistor Networks, Design of Resistive. Farhi, Groves...............................246–53
Temperature Reporting Over a Very-High-Frequency Radio System, Automatic Precipitation and. Warchol, Sachs, Lorentson.................77–80
10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer, A 30-Volt. Jackson.........518–22
Tentative Recommendation or the Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters. Gaines, Fischer.......... ....387–93; disc. 393
Testing of the Dynamic Performance of Analog-to-Digital Converters, Tentative Recommendation for the Definition and. Gaines, Fischer........ ....387–93; disc. 393
Tests Using Simulated Transmission Impairments, Data System. Andrews.......................590–6
Theory, E-Algebras in Switching. Semon.......265–9
Theory for Linear Time-Variable Systems, Signal Flow Graph. Stear, Stubberud...........695–701
Theory of a Single-Ring Circular Antenna Array, The. Hickman, Neff, Tillman..................110–15
Thermistor Networks, Design of Resistive Temperature Compensation by Single and Multiple. Farhi, Groves...............................246–53
Thermoelectric Generator, Electric Analog of Transport Phenomena in a. Talaat............427–31
Thermoelectric Generator, The Initial Transient Behavior of a. Stremler, Gray..............367–72
30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer, A. Jackson............518–22
Threshold Building Blocks for Variable-Radix Adders and Other Logic, Analog and. Sklansky...... 289–95
Time Saving and Better Measurement in the Speech Band, The Level Tracer, an Instrument for. Niedereder, Sanders......................674–82
Time-Variable Systems, Signal Flow Graph Theory for Linear. Stear, Stubberud..............695–701
Tone Disabler for Bell System 1A Echo Suppressors, A. Bugbee...........................596–600
Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band, The Level. Niedereder, Sanders......................674–82
Transient Behavior of a Thermoelectric Generator, The Initial. Stremler, Gray..................367–72
Transistor Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity. Mulligan, Shamis.................................335–40
Transistor Converter Circuits by Means of R-C Network, Control of Frequency and Phase Displacement in. Kadri.........................43–8
Transistor Feedback Amplifiers, Amplitude Distortion in. Mulligan..............................326–35
Transistor-Magnetic Inverters, Self-Locking Polyphase. Campling, Bennett........................26–33
Transistor Morse to Teleprinter Code Converter. Cunniff, Theall...........................178–81
Transistor Oscillator with Current Transformer, A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output. Jackson...........518–22
Transistorized Short-Haul Carrier System, Type T-12S-Tr Ishihara, Yamamoto, Fujita, Murakami 193–9
Transistorized 600-Channel Multiplex System, A Completely. Sparks........................121–6
Transfer Functions, Analysis of Conventional Single-Ended Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of. Nahi.... 394–400
Transformer, A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current. Jackson.........518–22
Transformers, Ultrahigh-Temperature Electronic:
  Part I. Rippin, Harms, Walters.............. ....295–302; disc. 302
  Part II. Walters.........................302–09
Transmission, A Buffer Store for Data. Baker, Froehlich 566–70
Transmission Impairments, Data System Tests Using Simulated. Andrews.....................590–6
Transmission Network of an Electronic Crosspoint PABX. Kowalik........................491–6
Transmission Performance, Telephone Circuit and Equipment Impedances as They Affect. Radue.. 126–34
Transmission System for Project EHV, Data Gathering and. Abetti, Caverly, Davis, Whepley.....272–81
Transmission System for Telephone Circuits, The Design of an Error-Free Data. Reiffen, Schmidt, Yudkin...................................224–31
Transmission Through Noisy Media—A Systems Approach, Reliable Data. Melas............501–04
Transmitting Station Parameters, Automatic Logging of AM, FM, and TV. Ehrenberg..........199–204
Transport Phenomena in a Thermoelectric Generator, Electric Analog of. Talaat...............427–31
Trigger Circuits, Design of Diode-Switch High-Sensitivity. Vojinovic.......................458–61
Tropospheric Scatter, Digital. Poppe, Vadasz....... 669–74
Trunk Carrier System—Applications, The 81-A Exchange. Cool, Walker..................617–23
Trunk Carrier System—Repeaters and Regulation, The 81-A Exchange. Gilmore, Griffith........606–10
Trunk Carrier System—Sigualing, The 81-A Exchange. Tuck, Melvin...........................610–17
Trunk Carrier System, The 81-A Exchange. Ferguson, Harp...............................601–06
Truth Function, An Algorithm for Determining Minimal Normal Forms of an Incomplete. Mott...73–6
Tube with Grid for Control of Exposures in Cineradiography, X-Ray. Rogers.................713–16
Tungsten Lamp Loads, Problems Encountered in Applying the Silicon-Controlled Rectifier to Control. Daugherty................................400–02
Tunnel Diode D-C Power Converter. Storm, Shattuck.................................347–52
Turn Gain of Ordinary Saturable Reactors, An Experimental Study on the Ampere-. Geyger......173–8
24-Bit Checked Binary Adder, A 4-Megacycle. Homan 443–50
20,000-Ampere Mechanical Switch, A. Newgard..96–8
Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines. Kelly, Mercurio, Willard................585–90
Two-Way Radio Channel for Maintenance of Telephone Company Services, Design of a. MacDonald.... 639–42
Type T-12S-Tr Transistorized Short-Haul Carrier System. Ishihara, Yamamoto, Fujita, Murakami 193–9

U

Ultrahigh-Temperature Electronic Transformers:
  Part I. Rippin, Harms, Walters.............. ....295–302; disc. 302
  Part II. Walters.........................302–09
Uniformity and Deviations Control in Polyethylene-Insulated Multipaired Cables, Electrical. Roberts. ....570–6; disc. 576
Unit for High-Frequency Vehicular Whip, Antenna Matching. Gruber, Seward..............99–104

V

Value of Direct-Voltage Ripple, Effective. Diebold.. ....61–6; disc. 66
Variable-Radix Adders and Other Logic, Analog and Threshold Building Blocks for. Sklansky...289–95
Vehicular Whip, Antenna Matching Unit for High-Frequency. Gruber, Seward..............99–104
Very-High-Frequency Radio System, Automatic Precipitation and Temperature Reporting Over a. Warchol, Sachs, Lorentson..................77–80
Very-High-Frequency Radio Systems, Remote Outage Location by Means of. Cooper.............635–9
Voice-Frequency Telephone Lines, Pole-Zero Techniques Applied to. Fleming................482–7
Voltage Regulator, The Design of the Generator. Van Emden...................................438–43
Voltmeters, A General Description of D-C Digital. Stansbury..............................465–70

W

Whip, Antenna Matching Unit for High-Frequency Vehicular. Gruber, Seward..............99–104
Wind of the Corona Discharge, Movement of Air in the Electric. Robinson......................143–50
(Wire Communication Systems) A Completely Transistorized 600-Channel Multiplex System. Sparks.. 121–6
(Wire Communication Systems) A Speakerphone System for Large Conference Rooms. Huggler.... ....237–40; disc. 240
(Wire Communication Systems) A Tone Disabler for Bell System 1A Echo Suppressors. Bugbee...... 596–600
(Wire Communication Systems) Automatic Dialing and Electronic Selection on Railroad Message and Dispatcher Circuits. Jorgensen..............204–08
(Wire Communication Systems) Effects of Cable Irregularities on the Operation of Telephone Repeaters. McCaron.......................217–14
(Wire Communication Systems) Electrical Uniformity and Deviations Control in Polyethylene-Insulated Multipaired Cables. Roberts.....570–6; disc. 576
(Wire Communication Systems) Pole-Zero Techniques Applied to Voice-Frequency Telephone Lines. Fleming................................482–6
(Wire Communication Systems) Synthesis of Networks Simulating Impedance of Loaded Cable Facilities. DeMonte..............................623–7
(Wire Communication Systems) Telephone Circuit and Equipment Impedances as They Affect Transmission Performance. Radue..................126–34
(Wire Communication Systems) The 81-A Exchange Trunk Carrier System. Ferguson, Harp....601–06
(Wire Communication Systems) The 81-A Exchange Trunk Carrier System—Applications. Cool, Walker.................................617–23
(Wire Communication Systems) The 81-A Exchange Trunk Carrier System—Repeaters and Regulation. Gilmore, Griffith.........................606–10
(Wire Communication Systems) The 81-A Exchange Trunk Carrier System—Signaling. Tuck, Melvin.. 610–17
(Wire Communication Systems) The Level Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band. Niedereder, Sanders.. 674–82
(Wire Communication Systems) The Subscriber Loop Bridge-Lifting Problem and Some Methods of Solution. Hochgraf.....................240–5
(Wire Communication Systems) Type T-12S-Tr Transistorized Short-Haul Carrier System. Ishihara, Yamamoto, Fugita, Murakami.............193–9
Wiring Diagrams, Computer Program for Preparing. Kirby, Rosenthal......................509–13
Wiring Process for Miniaturized Communications Electronics, A Stitched-. Coffin, Johnson.......... ....682–6; disc. 686

X

X-Ray Tube with Grid for Control of Exposures in Cineradiography. Rogers................713–16

Z

Zero Techniques Applied to Voice-Frequency Telephone Lines, Pole-. Fleming......................482–6

2. Author Index

Abetti, P. A.; D. W. Caverly, D. D. Davis, E. A. Whepley. Data Gathering and Transmission System for Project EHV. (60-1407)....272–81
Andrews, F. T., Jr. Data System Tests Using Simulated Transmission Impairments. (61-1012)..........590–6

B

Baker, W. E.; F. E. Froehlich. A Buffer Store for Data Transmission. (61-1020)....566–70
Baldwin, J. A., Jr. A Magnetic Device for High-Speed Sensing of Small Currents. (T-121)....1–3
Baldwin, M. E.; R. W. Moss, C. F. Kooi. Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites. (60-1015)....362–7
Baker, R. C.; R. M. Brownell. Detailed Measurements of Slow Magnetization Processes in Tape-Wound Cores. (61-129)....402–12
Barker, R. C.; A. J. Gruodis. A Saturable-Core Modulation Integrator. (T-121)....13–17
Beecher, A. E.; L. A. Gould. Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System. (60-980)....413–23
Bennett, J. A.; C. H. R. Campling. Self-Locking Polyphase Transistor-Magnetic Inverters. (T-121)....26–33
Bisson, D. K. Disc....191
Biswas, N. N. A Proposed Electronic Method for High-Speed Teleprinters. (61-29)....115–17
Bonnema, T.; G. R. Slemon. A Controllable Low-Ripple Constant-Current Source. (T-121)....3–8
Brand, S.; C. W. Carter. A 1,650-Bit-Per-Second Data System for Use Over the Switched Telephone Network. (61-1063)....652–61
Bromberg, Nathan S. Maximization and Minimization of Complicated Multivariable Functions. (61–803)....725–30; disc. 730
Bronikowski, R. J. Disc....288
Brothman, A.; E. H. Brothman, R. D. Reiser. Digital Telemetering Techniques. (61–17)....81–90
Brothman, E. H.; A. Brothman, R. D. Reiser. Digital Telemetering Techniques. (61–17)....81–90
Brownell, R. M.; R. C. Barker. Detailed Measurements of Slow Magnetization Processes in Tape-Wound Cores. (61–129)....402–12
Bugbee, L. F., Jr. A Tone Disabler for Bell System 1A Suppressors. (61–1069)....596–600
Burton, H. O.; W. R. Cowell. Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel. (61–1023)....577–85

C

Campling, C. H. R.; J. A. Bennett. Self-Locking Polyphase Transistor-Magnetic Inverters. (T-121)....26–33
Carter, C. W.; S. Brand. A 1,650-Bit-Per-Second Data System for Use Over the Switched Telephone Network. (61–1063)....652–61
Caverly, D. W.; P. A. Abetti, D. D. Davis, E. A. Whepley. Data Gathering and Transmission System for Project EHV. (60–1407)....272–81
Chapman, A. G. Disc....576
Clement, P. R.; W. C. Johnson. Motivation Through Challenge. (61–735)....554–6
Coffin, John M.; L. B. Johnson. A Stitched-Wiring Process for Miniaturized Communications Electronics. (61–1046)....682–6; disc. 686
Collins, C. A.; A. D. Williams. Noise and Intermodulation Problems in Multichannel Closed-Circuit Television Systems. (61–181)....486–91
Cool, L. R.; R. G. Walker. The 81-A Exchange Trunk Carrier System—Applications. (61–777)...617–23
Cooper, Martin. Remote Outage Location by Means of Very-High-Frequency Radio Systems. (61–920)....635–9
Cowell, W. R.; H. O. Burton. Computer Simulation of the Use of Group Codes with Retransmission on a Gilbert Burst Channel. (61–1023)....577–85
Craig, S.; G. W. Penney. Pulse Discharges Preceding Sparkover at Low Voltage Gradients. (61–91)....156–62
Cunniff, J. F.; C. E. Theall, Jr. Transistor Morse to Teleprinter Code Converter. (61–98)....178–81
Curl, F. G. A Comparison of Computers. (61–725)....542–6; disc. 547

D

Daugherty, C. S. Problems Encountered in Applying the Silicon-Controlled Rectifier to Control Tungsten Lamp Loads. (61–182)....400–02
Davis, D. D.; P. A. Abetti, D. W. Caverly, E. A. Whepley. Data Gathering and Transmission System for Project EHV. (60–1407)....272–81
DeCicco, J.; J. G. Van Bosse. Features of an Electronic Crosspoint PABX. (61–827)....471–4
DeMonte, R. W. Synthesis of Networks Simulating Impedance of Loaded Cable Facilities. (61–945)....623–7
Della Torre, E.; P. Dentella. The Determination of the Hysteresis Curve for Thick Tape Cores. (60–200)....359–62
Dentella, P.; E. Della Torre. The Determination of the Hysteresis Curve for Thick Tape Cores. (60–200)....359–62
Dickinson, M. M. Disc....192
Diebold, E. J. Effective Value of Direct-Voltage Ripple. (60–211)....61–6; disc. 67
Dortort, I. K. Disc....61
Drenning, J. W.; J. B. Thomas, H. T. Williams. Determination and Simulation of the Equivalent Circuits of Electrostatic Precipitators. (61–92)....315–19; disc. 319

E

Ehrenberg, Gustave. Automatic Logging of AM, FM, and TV Transmitting Station Parameters. (60–60)....199–204
Endsley, David L.; W. W. Grannemann, T. Summers. A Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion. (61–1)....150–2
Ewing, Ann C.; J. P. Roth, E. G. Wagner. Algorithms for Logical Design. (59–1191)....450–8

F

Farhi, Sivy; S. Groves. Design of Resistive Temperature Compensation by Single and Multiple Thermistor Networks. (61–189)....246–53
Fay, J. G.; R. J. Metz. Magnetics in Doppler Signal Data Extraction. (T-121)....33–43
Ferguson, M. E.; M. C. Harp. The 81-A Exchange Trunk Carrier System. (61–774)....601–06
Fischer, Peter P.; W. M. Gaines. Tentative Recommendation for the Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters. (61–15)....387–93; disc. 393
Fleming, Paul, Jr. Pole-Zero Techniques Applied to Voice-Frequency Telephone Lines. (61–113)....482–6
Flowers, T. H.; G. Forshaw. The British JERC Development of Electronic Switching. (61–75)....208–17
Forshaw, G.; T. H. Flowers. The British JERC Development of Electronic Switching. (61–75)....208–17
Friedlaender, F. J.; I. Leliakov. An Improved Model for Flux Reversal in Ni-Fe Cores. (T-121)....23–6
Friedlaender, F. J.; I. P. Leliakov. Initiation of Flux Reversal in Magnetic-Amplifier Cores. (60–874)....269–72
Froehlich, F. E.; W. E. Baker. A Buffer Store for Data Transmission. (61–1020)....566–70
Fujita, S.; O. Ishihara, Y. Yamamoto, R. Murakami. Type T-12S-Tr Transistorized Short-Haul Carrier System. (61–112)....193–9

G

Gaines, Warren M.; P. P. Fischer. Tentative Recommendation for the Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters. (61–15)....387–93; disc. 393
Geyger, William A. An Experimental Study on the Ampere-Turn Gain of Ordinary Saturable Reactors. (61–35)....173–8
Geyger, William A. Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of Ink Recorders. (T-121)....8–13
Gilmore, E. A.; C. G. Griffith. The 81-A Exchange Trunk Carrier System—Repeaters and Regulation. (61–775)....606–10
Gould, L. A.; A. E. Beecher. Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System. (60–980)....413–23
Grannemann, W. W.; D. L. Endsley, T. Summers. A Low-Level Linear Rundown Circuit for Pulse Height to Pulse-Width Conversion. (61–1)....150–2
Gray, P. E.; F. G. Stremler. The Initial Transient Behavior of a Thermoelectric Generator. (61–225)....367–72
Green, W. B. A Fatigue-Free Silicon Device Structure. (61–24)....186–91; disc. 191
Greene, W. J.; E. A. Saibel. Analog Computer Stabilization of High-Speed Magnetic Amplifiers. (T-121)....17–22
Greenwood, Allan, Disc....172
Griffith, C. G.; E. A. Gilmore. The 81-A Exchange Trunk Carrier System—Repeaters and Regulation. (61–775)....606–10
Gross, Eric T. B. Disc....553
Groves, Stanley; S. Farhi. Design of Resistive Temperature Compensation by Single and Multiple Thermistor Networks. (61–189)....246–53
Gruber, J. R.; G. J. Seward. Antenna Matching Unit for High-Frequency Vehicular Whip. (61–47)....99–104
Gruodis, A. J.; R. C. Barker. A Saturable-Core Modulation Integrator. (T-121)....13–17

H

Hall, Frank B. Boolean Prime Implicants by the Binary Sieve Method. (60–1283A)....709–13
Harms, H. B.; J. F. Rippin, G. Walters. Ultrahigh-Temperature Electronic Transformers:
  Part I. (60–231)....295–302; disc. 302
  (Part II by Walters)
Harp, M. C.; M. E. Ferguson. The 81-A Exchange Trunk Carrier System. (61–774)....601–06
Henches, E. W. A New Billing and Collection System Using Teletypewriter Techniques. (61–1029)....560–6
Hermach, F. L. An Analysis of Errors in the Calibration of Electric Instruments. (61–63)....90–5; disc. 95
Heumann, G. W. Disc....287
Hickman, C. E.; H. P. Neff, J. D. Tillman. The Theory of a Single-Ring Circular Antenna Array. (61–61)....110–15
Hochgraf, Lester. The Subscriber Loop Bridge-Lifting Problem and Some Methods of Solution. (61–164)....240–5
Hollander, Gerhard L. Disc....192, 550
Homan, M. E. A 4-Megacycle 24-Bit Checked Binary Adder. (61–193)....443–50
Huggler, Luther R. A Speakerphone System for Large Conference Rooms. (60–1250)....237–40; disc. 240
Hurwitz, H., Jr.; G. W. Sutton, H. Poritsky. Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops. (61–1078)....687–95

I

Ishihara, O.; Y. Yamamoto, S. Fujita, R. Murakami. Type T-12S-Tr Transistorized Short-Haul Carrier System. (61–112)....193–9

J

Jackson, S. P. A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer. (61–717)....518–22
Jacobs, J. E. Technical Solutions to the Problem of Reducing Patient Dosage. (69–236)....721–5
Johnson, L. B.; J. M. Coffin. A Stitched-Wiring Process for Miniaturized Communications Electronics. (61–1046)....682–6; disc. 686
Johnson, W. C.; P. R. Clement. Motivation Through Challenge. (61–735)....554–6
Jones, Sheldon. On the Elimination of Null in Modulating and Demodulating Devices. (61–25)....135–9; disc. 139
Jorgensen, Otto A. Automatic Dialing and Electronic Selection on Railroad Message and Dispatcher Circuits. (61–115)....204–08

K

Kadri, F. V. Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C Networks. (T-121)....43–8
Kaufmann, W. Disc....287
Kelly, J.; J. Mercurio, D. Willard. Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines. (61–1065)....585–90
Kirby, D. B.; C. W. Rosenthal. Computer Program for Preparing Wiring Diagrams. (60–1007)....509–13
Kooi, C. F.; R. W. Moss, M. E. Baldwin. Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites. (60–1015)....362–7
Kowalik, R. F. Transmission Network of an Electronic Crosspoint PABX. (61–825)....491–6

L

Laaspere, T. A New Method Space Exploration by Radar, Using Incoherent Scattering. (61–1048)....557–60

Leliakov, I.; F. J. Friedlaender. An Improved Model for Flux Reversal in Ni-Fe Cores. (T-121)....23-6
Leliakov, I. P.; F. J. Friedlaender. Initiation of Flux Reversal in Magnetic-Amplifier Cores. (60-874)....269-72
Lentz, R. K. Santa Barbara County 6,000-Mc Microwave System. (61-1113)....627-34
LePage, W. R. Disc....553
Lerner, Robert M. Modulation and Signal Selection for Digital Data Systems. (61-1096)....661-9
Lerstrup, Karl. Disc....286
Levine, S. Large-Scale On-Line Data Processing Systems. (61-732)....505-09
Linke, Simpson. Educating Electrical Engineers for Professional Careers (61-733)....551-3; disc. 553
Lorentson, H. S.; E. J. Warchol, M. S. Sachs. Automatic Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System. (60-992)....77-80

M

MacDonald, D. L. Design of a 2-Way Radio Channel for Maintenance of Telephone Company Services. (61-927)....639-42
MacIndoe, R. C.; J. C. Nycz, V. S. Thomander, D. F. Roerty. Taut Band Suspension Switchboard Instruments. (61-65)....258-64
Mathes, K. N.; E. J. McGowan. Surface Electrical Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test. (61-21)....281-6; disc. 286
McCaron, S. I. Effects of Cable Irregularities on the Operation of Telephone Repeaters. (61-118)....217-14
McGowan, E. J.; K. N. Mathes. Surface Electrical Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test. (61-21)....281-6; disc. 286
McMurray, W.; D. P. Shattuck. A Silicon-Controlled Rectifier Inverter with Improved Commutation. (61-718)....531-42
Meade, R. M. Forcing Circuitry: Sequential Building Blocks for Logical Design. (61-814)....522-31
Melas, C. M. Reliable Data Transmission Through Noisy Media—A Systems Approach. (61-822)....501-04
Melvin, D. K.; E. F. Tuck. The 81-A Exchange Trunk Carrier System—Signaling. (61-776)....610-17
Mercurio, J.; J. Kelly, D. Willard. Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines. (61-1065)....585-90
Metz, R. J.; J. G. Fay. Magnetics in Doppler Signal Data Extraction. (T-121)....33-43
Metzner, J. J.; K. C. Morgan. Coded Feedback Comunication Systems. (61-1050)....643-7
Metzner, J. J.; K. C. Morgan. Disc....652
Mitrovic, D. M. Automatic Plotting of Characteristic Curves and Analog Solution of Algebraic Equations. (61-1053)....701-03; disc. 703
Miyata, J. J. Optical Read-out of Digital Magnetic Recording. (T-121)....53-8
Morgan, K. C.; J. J. Metzner. Coded Feedback Communication Systems. (61-1050)....643-7
Morgan, K. C.; J. J. Metzner. Disc....652
Morgan, Ray E. A New Magnetic-Controlled Rectifier Power Amplifier with a Saturable Reactor Controlling On Time. (T-121)....152-5
Moss, R. W.; C. F. Kooi, M. E. Baldwin. Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites. (60-1015)....362-7
Mott, Thomas H., Jr. An Algorithm for Determining Minimal Normal Forms of an Incomplete Truth Function. (59-1123)....73-6
Moyer, D. F.; G. D. Talbott. Instrumentation for the Diagnosis of Coronary-Artery Disease. (60-100)....717-21
Mulligan, J. H., Jr. Amplitude Distortion in Transistor Feedback Amplifiers. (61-97)....326-35
Mulligan, J. H., Jr.; S. S. Shamis. Transistor Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity. (61-96)....335-40
Murakami, R.; O. Ishihara, Y. Yamamoto, S. Fujita. Type T-12S-Tr Transistorized Short-Haul Carrier System. (61-112)....193-9

N

Nahi, N. E. Analysis of Conventional Single-Ended Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of Transfer Functions. (61-148)....394-400
Neff, H. P.; C. E. Hickman, J. D. Tillman. Theory of a Single-Ring Circular Antenna Array. (61-61)....110-15
Newgard, Peter M. A 20,000-Ampere Mechanical Switch. (61-134)....96-8
Niedereder, Martin; B. J. Sanders. Jr. The Level Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band. (61-1072)....674-82
Nycz, J. C.; V. S. Thomander, R. C. MacIndoe, D. F. Roerty. Taut Band Suspension Switchboard Instruments. (61-65)....258-64

P

Pande, H. C. Linear Analysis of the Electromechanical Pulser. (61-100)....309-15
Peek, R. L., Jr. Magnetization and Pull Characteristics of Mating Magnetic Reeds. (60-1282)....372-80
Penney, Gaylord W. Disc....319
Penney, G. W.; S. Craig. Pulse Discharges Preceding Sparkover at Low-Voltage Gradients. (61-91)....156-62
Penney, G. W.; P. L. Smith. The Charging of Nonspherical Particles in a Corona Discharge. (61-93)....340-6
Perkinpine, R. W. Disc....686
Peterson, W. W. Binary Controls for Error Control. (61-1097)....648-52; disc. 652
Polansky, R. B. Minimization of Multiple-Output Switching Circuits. (60-1223)....67-73
Poppe, J. R.; A. J. Vadasz. Digital Tropospheric Scatter. (61-1112)....669-74
Poritsky, H.; G. W. Sutton, H. Hurwitz, Jr. Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops. (61-1078)....687-95
Prywes, Noah S. Relaxation Oscillations at Multiple Frequencies of Even Subharmonics. (T-121)....48-52

R

Radue, Paul F. Telephone Circuit and Equipment Impedances as They Affect Transmission Performance. (61-117)....126-34
Redington, T. R. A Low-Cost Number Identifier for Small Telephone Systems. (61-69)....117-21
Reiffen, Barney; W. G. Schmidt, H. L. Yudkin. The Design of an Error-Free Data Transmission System for Telephone Circuits. (61-101)....224-31
Reiser, R. D.; A. Brothman, E. H. Brothman. Digital Telemetering Techniques. (61-17)....81-90
Rippin, J. F.; H. B. Harms, G. Walters. Ultrahigh-Temperature Electronic Transformers:
  Part I. (60-231)....295-302; disc. 302
  (Part II by Walters)
Rivenes, R. A. Disc....240
Roberts, W. L. Electrical Uniformity and Deviations Control in Polyethylene-Insulated Multipaired Cables. (61-779)....570-6; disc. 576
Robinson, Myron. Movement of Air in the Electric Wind of the Corona Discharge. (61-90)....143-50
Roerty, D. F.; J. C. Nycz, V. S. Thomander, R. C. MacIndoe. Taut Band Suspension Switchboard Instruments. (61-65)....258-64
Rogers, T. H. X-Ray Tube with Grid for Control of Exposures in Cineradiography. (61-34)....713-16
Rosenthal, C. W.; D. B. Kirby. Computer Program for Preparing Wiring Diagrams. (60-1007)....509-13
Roth, J. Paul; A. C. Ewing, E. G. Wagner. Algorithms for Logical Design. (59-1191)....450-8
Rowe, Ian H.; G. R. Slemon. On the Process of Flux Reversal in Multiaperture Ferrite Cores. (61-138)....431-8

S

Sachs, M. S.; E. J. Warchol, H. S. Lorentson. Automatic Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System. (60-992)....77-80
Sager, Guenter. A Magnetically Regulated D-C to D-C Converter Power Supply. (61-150)....513-18
Saibel, E. A.; W. J. Greene. Analog Computer Stabilization of High-Speed Magnetic Amplifiers. (T-121)....17-22
Samelson, Lincoln R. Disc....554
Sanders, Benjamin T., Jr.; M. Nierdereder. The Level Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band. (61-1072)....674-82
Sanders, R. P. Logical Control of an Electronic Crosspoint PABX. (61-826)....496-501
Saucedo, R.; T. W. Sze. Analog Simulation of Digital Computer Programs. (61-1060)....703-09
Scherer, E. H. An Error-Correction Code for Quaternary Data Transmission. (61-170)....231-6
Schindler, M. J. Design of High Coercivity Permanent Magnets, Exposed to External Fields. (61-194)....423-7
Schlereth, F. H. A Method of Calculating Gain and Phase Margins from Pole-Zero Diagrams. (60-1290)....139-43
Schmidt, W. G.; B. Reiffen, H. L. Yudkin. The Design of an Error-Free Data Transmission System for Telephone Circuits. (61-101)....224-31
Schubert, E. J. Logical Design by Regression—Matrix Logic VII. (60-1229)....380-3
Schubert, Ernest J. Disc....546
Semon, Warren. E-Algebras in Switching Theory. (60-1266)....265-9
Sendzuk, G. T. Results of Simulation Comparison of Control Computers. (61-724)....547-50; disc. 550
Seward, G. J.; J. R. Gruber. Antenna Matching Unit for High-Frequency Vehicular Whip. (61-47)....99-104
Shamis, S. S.; J. H. Mulligan, Jr. Transistor Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity. (61-96)....335-46
Shattuck, D. P.; W. McMurray. A Silicon-Controlled Rectifier Inverter with Improved Communication. (61-718)....531-42
Shattuck, D. P.; H. F. Storm. Tunnel Diode D-C Power Converter. (61-149)....347-52
Shea, P. D. Centrex Service: A New Design for Customer Group Telephone Service in the Modern Business Community. (61-722)....474-82
Shiva, S. G. S. A Note on the Approximate Representation of a Curve with Linear Segments. (61-804)....461-3
Sklansky, J. Analog and Threshold Building Blocks for Variable-Radix Adders and Other Logic. (60-805)....289-95
Slemon, G. R.; T. Bonnema. A Controllable Low-Ripple Constant Current Source. (T-121)....3-8
Slemon, Gordon R.; I. H. Rowe. On the Process of Flux Reversal in Multiaperature Ferrite Cores. (61-138)....431-8
Smith, P. L. Disc....319
Smith, P. L.; G. W. Penney. The Charging of Nonspherical Particles in a Corona Discharge. (61-93)....340-6
Sommerman, G. M. L. Disc....287
Somos, Istvan. Commutation and Destructive Oscillation in Diode Circuits. (61-89)..162-72; disc. 172
Somos, I. Switching Characteristics of Silicon Power-Controlled Rectifiers:
  Part I. (60-227)....320-6
Sparks, D. W. A Completely Transistorized 600-Channel Multiplex System. (61-111)....121-6
Stansbury, Carroll. A General Description of D-C Digital Voltmeters. (61-721)....465-70
Stear, E. B.; A. R. Stubberud. Signal Flow Graph Theory for Linear Time-Variable Systems. (61-1051)....695-701
Steinbach, L. J. Disc....302
Storm, H. F.; D. P. Shattuck. Tunnel Diode D-C Power Converter. (61-149)....347-52
Stremler, F. G.; P. E. Gray. The Initial Transient Behavior of a Thermoelectric Generator. (61-225)....367-72
Stricker, S.; A. A. Wulkan. A Pulse-Position Modulalator-Type Magnetometer. (61-107)....253-8
Stubberud, A. R.; E. B. Stear. Signal Flow Graph Theory for Linear Time-Variable Systems. (61-1051)....695-701
Summers, Thomas; D. L. Endsley, W. W. Grannemann. A Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion. (61-1)....150-2
Sutton, G. W.; H. Hurwitz, Jr., H. Poritsky. Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops. (61-1078)....687-95
Sze, T. W.; R. Saucedo. Analog Simulation of Digital Computer Programs. (61-1060)....703-0

T

Talaat, M. E. Electric Analog of Transport Phenomena in a Thermoelectric Generator. (61-188)....427-31
Talbott, G. D.; D. F. Moyer. Instrumentation for the Diagnosis of Coronary-Artery Disease. (60-100)....717-21
Teager, Herbert M. Disc....730
Temes, G. C. The Synthesis of General Parameter Insertion Loss Filters Using a Digital Computer (61-5)....181-6
Thaler, George J. Disc....703
Theall, C. E., Jr.; J. F. Cunniff. Transistor Morse to Teleprinter Code Converter. (61-98)....178-81
Thomander, V. S.; J. C. Nycz, R. C. MacIndoe, D. F. Roerty. Taut Band Suspension Switchboard Instruments. (61-65)....258-64
Thomas, J. B.; J. W. Drenning, H. T. Williams. Determination and Simulation of the Equivalent Circuits of Electrostatic Precipitators. (61-92)....315-19; disc. 319

Tillman, J. D.; C. E. Hickman, H. P. Neff. The Theory of a Single-Ring Circular Antenna Array. (61–61)....110–15
Tuck, E. F.; D. K. Melvin. The 81-A Exchange Trunk Carrier System—Signaling. (61–776)....610–17

V

Vadasz, A. J.; J. R. Poppe. Digital Tropospheric Scatter. (61–1112)....669–74
Van Bosse, J. G.; J. DeCicco. Features of an Electronic Crosspoint PABX. (61–827)....471–4
Vanderleck, J. M. Disc....95
Van Emden, Bernard M. The Design of the Generator Voltage Regulator. (60–1413)....438–43
Vojinovic, Mirko M. Design of Diode-Switch High-Sensitivity Trigger Circuits. (61–99)....458–61

W

Wagner, Eric G.; A. C. Ewing, J. P. Roth. Algorithms for Logical Design. (59–1191)....450–8
Walker, R. G.; L. R. Cool. The 81-A Exchange Trunk Carrier System—Applications. (61–777)....617–23
Walters, Glenn. Ultrahigh-Temperature Electronic Transformers:
  Part II. (60–230)....302–09
  (Part I by Rippin, Harms, Walters)
Walters, Glenn; J. F. Rippin, H. B. Harms. Ultrahigh-Temperature Electronic Transformers:
  Part I. (61–231)....295–302; disc. 302
  (Part II by Walters)
Walton, Charles A. Disc....393
Warchol, E. J.; M. S. Sachs, H. S. Lorentson. Automatic Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System. (60–992)....77–80
Weeks, R. W. Rotating Raster Character Recognition System. (60–1005)....353–9
Weiss, Gerald. Disc....139
Whepley, E. A.; P. A. Abetti, D. W. Caverly, D. D. Davis. Data Gathering and Transmission System for Project EHV. (60–1407)....272–81
Willard, D.; J. Kelly, J. Mercurio. Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines. (61–1065)....585–90
Williams, A. D.; C. A. Collins. Noise and Intermodulation Problems in Multichannel Closed-Circuit Television Systems. (61–181)....486–91
Williams, H. T.; J. B. Thomas, J. W. Drenning. Determination and Simulation of the Equivalent Circuits of Electrostatic Precipitators. (61–92)....315–19; disc. 319
Wolf, Alfred A. General Recurrence Relations in the Analysis of Nonlinear Systems. (60–193)....383–7
Wulkan, A. A.; S. Stricker. A Pulse-Position Modulator-Type Magnetometer. (61–107)....253–8
Wusteney, Herbert J. Telegraph Distortion and Distortion Measuring. (61–60)....104–09

Y

Yamamoto, Y.; O. Ishihara, S. Fujita, R. Murakami. Type T-12S-Tr Transistorized Short-Haul Carrier System. (61–112)....193–9
Yudkin, Howard L.; B. Reiffen, W. G. Schmidt. The Design of an Error-Free Data Transmission System for Telephone Circuits. (61–101)....224–31

A Magnetic Device for High-Speed Sensing of Small Currents

J. A. BALDWIN, JR.
NONMEMBER AIEE

A MAGNETIC DEVICE, capable of sensing small currents, has been developed. A bar or stick of low-coercive-force square-loop ferrite forms the core of the device, and two holes pierce the stick, transverse to its long axis. The current to be detected flows in a solenoid, which surrounds the stick. The sensing of this current is accomplished by observing the flux switched around the holes under the action of a pulse of current flowing through them.

The material around the holes behaves like a pair of toroidal cores whose switchable flux may be controlled by the solenoid current. This flux may be made to be independent of solenoid currents below a certain critical value, I_1. It declines in an approximately linear manner between I_1 and another, higher current value, I_2; it remains small for currents above I_2. I_1 and I_2 may be independently varied by changing the core's dimensions.

This paper develops a theory which explains the observed behavior. Equations are derived which give the dependence of the output on the magnetic field generated by the solenoid in terms of the saturation induction and coercive force of the ferrite and of the core dimensions. Typical experimental data are presented and compared with the predictions of the theory.

In some systems it is necessary to monitor the status of a large number of slowly varying inputs by high-speed scanning. The device described in this paper, called the ferrod, was developed for this application. It is briefly described and its operating characteristics indicated, but the bulk of the paper is devoted to explaining the observed magnetic behavior.

Structure of the Device

If the input signals to be monitored exist as voltages, they may be impressed across solenoids. The field generated by a solenoid could then be used to saturate a toroid of some low-coercive-force square-loop ferrite material. The presence or absence of current in the solenoid would be detected by observing the voltage induced on a sensing winding, threading the toroid by the action of alternating-polarity current pulses passed through an interrogating winding, also threading the toroid.

The ferrod operates in essentially this way. The core (the word is used here in its original sense, meaning it is a piece of ferromagnetic material in electromagnetic interaction with one or more windings), however, instead of being a toroid, has the form of a stick of ferrite pierced by two holes as shown in Fig. 1. Both the sense and interrogate conductors pass out through one hole and back through the other. The solenoid, which is not shown, has the form of two coils connected in series aiding. One coil surrounds the upper portion of the core and the other the lower.

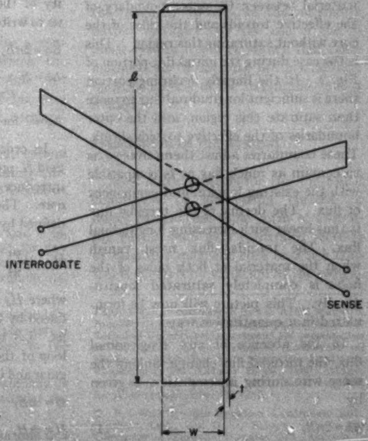


Fig. 1. A ferrod core with interrogate and sense conductors

Electrical Properties

In order to obtain experimental data which could be compared directly with the predictions of theory, measurements of the following sort were made. The core was immersed in the magnetic field of a solenoid which produced a substantially uniform field over the complete length of the core. A train of alternating-polarity current pulses, of sufficient magnitude to completely switch the material surrounding the holes, was passed through the interrogation winding. The voltage induced on the sense winding was integrated and the resulting flux change displayed on an oscilloscope. The remanent flux change during interrogation was measured. When this is plotted as a function of solenoid field, curves like that shown by points in Fig. 2 are obtained. There is an initial flat region in which the flux change does not respond to the solenoid field. This is followed by a region of linear decline. Finally, there is a tail in which the output falls to zero.

Theory of Operation

The portion of the core containing the holes is shown in Fig. 3. The core has a width w and thickness t. The holes have a diameter d and center-to-center spacing $r+d$. We suppose that the material which switches during interrogation is confined to the effective toroids shown as dashed circles in the figure.

In the absence of any magnetic field in the solenoid, a pulse of current through the interrogation winding will magnetize

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 15, 1960; made available for printing October 13, 1960.

J. A. Baldwin, Jr., is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

The author wishes to acknowledge the work of H. F. May, who, jointly with the author undertook exploratory development of ferrod structures, the basis of this paper.

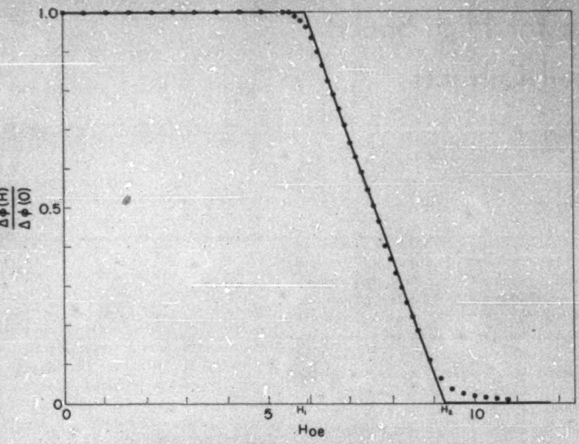


Fig. 2. Flux switched during interrogation versus longitudinal field for core of Table I. Points are experimental; solid line is theory

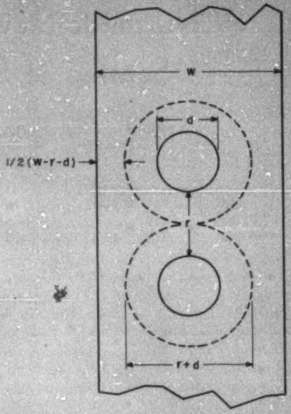


Fig. 3. Center portion of core enlarged to show effective toroids containing switching flux

one of the effective toroids in a clockwise and the other in a counterclockwise direction. If this pulse is followed by another of opposite polarity, the magnetization direction of the toroids will be reversed. The flux change picked up by the sense winding during this reversal will be twice the saturation induction of the material, times the area of minimum cross section between the holes.

A component of flux directed along the length of the core will not interfere with the switching of the effective toroids so long as that flux can pass through the material between the outer boundary of the effective toroids and the sides of the core without saturating this region. This is the case during the initial flat portion of Fig. 2. In the linearly declining portion there is sufficient longitudinal flux to more than saturate this region, and the outer boundaries of the effective toroids shrink. These boundaries adjust themselves so as to contain as much flux as is compatible with the existing longitudinal component of flux. The decline of the toroidal flux is thus linear with increasing longitudinal flux. The toroidal flux must vanish when the material at both sides of the holes is completely saturated longitudinally. This picture will now be formulated in a quantitative way.

In the absence of any longitudinal flux, the toroidal flux change, linking the sense wire during interrogation, is given by

$$\Delta\phi=2rtB_s \tag{1}$$

where $\Delta\phi$ is the flux switched between the holes; r is the minimum distance between the holes; t is the core thickness; and B_s is the saturation induction of the material. Fig. 3 shows that a maximum longitudinal flux ϕ_l given by

$$\phi_l=t(w-r-d)B_s \tag{2}$$

may exist without interfering with the toroidal flux. On the other hand, the toroidal flux vanishes when

$$\phi_l=t(w-d)B_s \tag{3}$$

These equations, together with the linearity of the decline of $\Delta\phi$ with ϕ_l, allow us to write

$$\Delta\phi=2rtB_s \text{ for } \phi_l\leq t(w-r-d)B_s \tag{4}$$

$$\Delta\phi=2t(w-d)B_s-2\phi_l \text{ for } t(w-r-d)B_s\leq\phi_l\leq t(w-d)B_s$$

$$\Delta\phi=0 \text{ for } t(w-d)B_s\leq\phi_e$$

In order to relate ϕ_l to the magnetic field H, produced by the solenoid, we must introduce the demagnetizing factor of the core. The demagnetizing factor D is defined by

$$H_d=D\frac{\phi_l}{wt} \tag{5}$$

where H_d is the demagnetizing field produced by the longitudinal induction, ϕ_l/wt. Let us suppose that the hysteresis loop of the material is perfectly rectangular and is bounded by the lines

$$B=\pm B_s$$

$$H=\pm H_c \tag{6}$$

where H_c is the coercive force. Then the longitudinal flux produced by the external field H will be such that

$$H=H_d+H_c \tag{7}$$

provided the core does not saturate. The required relation is then

$$\phi_l=(H-H_c)\frac{wt}{D} \tag{8}$$

A more accurate relationship may be obtained by the usual geometric construction on the appropriate actual hysteresis curve of the material.

Equations 4 and 8 suffice to derive the flux available for switching as a function of solenoid field. This is given subsequently, after first defining certain critical fields.

$$H_D=DB_s \tag{9}$$

$$H_1=H_c+\frac{w-r-d}{w}H_D$$

$$H_2=H_c+\frac{w-d}{w}H_D$$

H_D is introduced for convenience. It is the demagnetizing field a saturated core would experience. H_1 is the field at which $\Delta\phi$ first begins to decline; H_2 is the field at which it reaches zero. In terms of these fields, $\Delta\phi$ is given by

$$\Delta\phi=2rtB_s \text{ for } H\leq H_1 \tag{10}$$

$$\Delta\phi=2wtB_s\frac{H_2-H}{H_D}\text{ for } H_1\leq H\leq H_2$$

$$\Delta\phi=0 \text{ for } H_2\leq H$$

The holes may be moved laterally with-

Table I. Core Dimensions

$w = 0.12650$ inch
$r = 0.02740$ inch
$d = 0.05226$ inch
$t = 0.04920$ inch
$l = 1.5007$ inches
$B_s = 2{,}200$ gauss or unit of magnetic flux density
$H_c = 0.2$ oersted

out affecting these results as long as the effective toroids do not touch the sides of the core.

It remains to find an expression for the demagnetizing factor. We would expect the demagnetizing field in the center of the core not to differ greatly from that inside an ellipsoid whose axes were the core length, width, and thickness. Making the necessary changes in the equation for the demagnetizing factor of a slender ellipsoid,[1] we obtain

$$D = \frac{wt}{l^2}\left[\log_e\left(\frac{4l}{w+t}\right) - 1\right] \qquad (11)$$

Comparison with Experiment

The data of Fig. 2 were taken, using a core made from a low-coercive-force square-loop manganese-magnesium-cadmium ferrite. The core dimensions are shown in Table I.

The experimental points are so plotted that the flux change is unity in the absence of applied field. The measurement of flux switched as a function of external field was done as outlined above. Equations 9 and 11 give the critical fields for this core.

$$H_1 = 5.90 \text{ oersteds}; \; H_2 = 9.23 \text{ oersteds} \qquad (12)$$

Equation 10 is plotted as a solid line for comparison.

Discussion

By moving the holes closer together and farther away, one can vary the "knee" field, H_1, between zero and the "ankle" field, H_2. Thus, the decline of $\Delta\phi$ may be made to occur either gradually or abruptly.

The current sensitivity may be controlled—either through the design of the solenoid or through the demagnetizing factor.

The core described here is an ideal one. The holes have the same diameter and their centers lie on a line parallel to the core's axis. If these conditions do not obtain, the analysis is more complicated, but it can still be carried through. Inequality of the hole diameters leads to a more pronounced "tail" because the saturation of material beside the larger hole makes it difficult to get sufficient flux past to cut off the smaller. If the line through centers of the holes is not parallel to the core axis, the toroids present a greater area to the longitudinal flux than they otherwise would. For this reason the decline of $\Delta\phi$ begins sooner. The "knee" is, as a result, more rounded. If the core is made from material which does not have a rectangular hysteresis loop, the knee and tail regions extend over much wider ranges of H.

Fundamentally, the ferrod makes use of two interacting flux configurations. One of these has high remanence; the other, low. The latter controls the former while the reverse interaction is small. It is thus a magnetic amplifier. The fact that the control winding is on a bobbin rather than a toroid makes it economically attractive for many applications.

Reference

1. Demagnetizing Factors of the General Ellipsoid, J. A. Osborn. *Physical Review*, New York, N. Y., vol. 67, 1945, p. 351.

A Controllable Low-Ripple Constant-Current Source

T. BONNEMA
STUDENT MEMBER AIEE

G. R. SLEMON
MEMBER AIEE

Synopsis: This paper describes a controllable direct-current supply in which the current is essentially independent of the load and in which very low ripple is achieved at essentially all values of controlled current output. The system consists of a 3-phase saturable reactor with a high-impedance control circuit used in conjunction with uncontrolled rectifiers. An analysis of the system is presented and experimental results are given.

WHILE most d-c applications require a fixed or controllable voltage supply, there are a number of situations in which a controllable d-c supply having a current-source property would be useful. For example, such a current source might be used (1) for the torque control of d-c motors, (2) for the field excitation of d-c and synchronous machines, (3) for battery charging, and (4) for electroplating. In addition, there are some applications which require a controllable current source in which the ripple current must be very small. An example is the current supply for electromagnets in certain particle accelerators.

This paper describes a static system which converts 3-phase alternating current to direct current and has the following features:

1. The current supplied to a load is essentially independent of the load for any value of load resistance between zero and the value which results in maximum designed load voltage.
2. The load current is linearly dependent on the direct control current of the system over the control range from zero to the maximum designed value.
3. The ripple on the load current may be made inherently less than 1% without the use of a load filter.
4. The system has the control gain and time-constant characteristics of a standard saturable reactor.

H. F. Storm has described a single-phase saturable reactor-rectifier system which produces some of the above features, but has the limitations that the linear control characteristic and the ripple-free output are obtained only for the range up to 84% of the maximum output voltage.[1] By the use of the 3-phase system it will be shown that these features can be obtained up to a limiting value of 98.5% of the maximum output voltage.

Description of the System

Most static systems for the controlled conversion of alternating current to direct current use some form of switching device such as a controlled rectifier or saturable magnetic core in which control is effected by variation of the switch-

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted August 1, 1960; made available for printing September 21, 1960.

T. Bonnema is with Canadian Controllers Ltd. Scarboro, Ont., and G. R. Slemon is with the University of Toronto, Toronto, Ont., Canada.

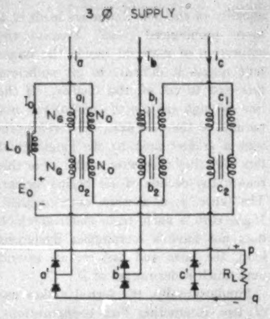


Fig. 1. Schematic diagram of controlled constant current system

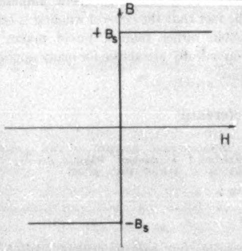


Fig. 2. Idealized B-H characteristic of core material

ing angle. The d-c output appears as a sequence of sections of the applied alternating voltage having a controlled average value. Filtering of this output to produce a low value of ripple may be difficult, particularly for low values of output voltage. An alternative approach to the production of a low-ripple direct current is to place in series with an uncontrolled rectifier a device which maintains the rectifier current at a constant controlled value. The basis for such a device is a 2-winding magnetic core which, in its unsaturated state, has essentially equal and opposite magnetomotive forces in its two windings. If the current in one winding is constrained to be constant, the current in the other winding will also be constant while the core is unsaturated.

Fig. 1 is a schematic diagram of the system to be discussed. A 3-phase supply is connected to a 3-phase full-wave bridge rectifier through windings of two magnetic cores in each of the three phases. Each magnetic core has two windings, a gate winding with N_G turns and a control winding with N_O turns. The gate windings of the two cores in each phase are connected in series aiding and the control windings in series opposition. The control windings of all six cores are then connected in series through an inductor L_O to the control voltage E_O as shown in Fig. 1. It has been shown[1] that the total induced voltage in the six series-connected control windings contains only sixth and multiples of the sixth harmonic of the supply frequency. These harmonic voltages are readily absorbed by the inductor L_O. Thus the control current I_O will be essentially free of harmonic components and will have a constant controllable value equal to the control voltage E_O divided by the total control circuit resistance R_O. Having obtained a constant control winding current, the action of the cores is to produce a constant, non-time-varying current in the load as will be shown in the following analysis.

Analysis of the System

For the theoretical analysis the following assumptions are made:

1. The **B-H** characteristic of the core material is approximated by the idealized characteristic of Fig. 2.
2. The leakage fluxes and resistances of the core windings are assumed to be negligible.
3. The rectifiers are assumed to have negligible resistance in the forward direction and negligible leakage current in the reverse direction.

First, relate the flux linkages of cores a_1 and a_2 in Fig. 1 to the phase current i_a and the control current I_O. Suppose each core has a cross-sectional area A. Then, from Fig. 2, the maximum value of the gate-winding flux linkage of a single core is

$$\lambda_m = N_G B_s A \tag{1}$$

Also, from Fig. 2, if a core is to be in an unsaturated condition its gate and control magnetomotive forces must be equal and opposite. From Fig. 1, core a_1 will be unsaturated only when

$$i_a = -\frac{N_O I_O}{N_G} \tag{2}$$

while core a_2 will be unsaturated only when

$$i_a = \frac{N_O I_O}{N_G} \tag{3}$$

Core a_1 will have a flux linkage of $-\lambda_m$ for values of i_a less than that of equation 2 and of λ_m for values of i_a greater than that of equation 2. A similar statement applies for core a_2 if i_a from equation 3 is used. The total flux linkages $\lambda_{aa'}$ of the branch aa' in Fig. 1 may therefore be related to the phase current i_a and the control current I_O by the characteristic of Fig. 3. Both cores are saturated in the interval

$$-\frac{N_O}{N_G} I_O < i_a < \frac{N_O}{N_G} I_O \tag{4}$$

and are able to absorb voltage by changing their total flux linkage $\lambda_{aa'}$ only at the values of i_a given in equations 2 and 3.

Operation with Short-Circuited Load

Fig. 4 shows the balanced supply voltages of the system. For convenience in analysis, all voltages of modes in Fig. 1 are measured with respect to the neutral point of this supply, except where noted otherwise by the use of a double subscript.

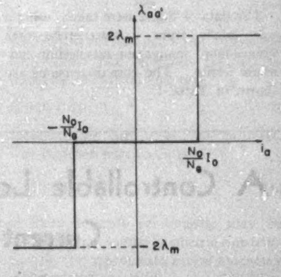


Fig. 3. Flux linkage $\lambda_{aa'}$ of cores in phase a as function of phase current i_a and control current I_O

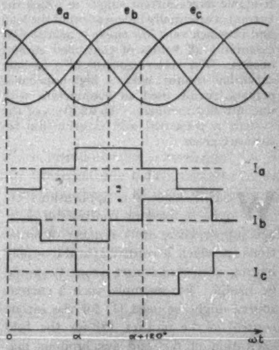


Fig. 4. Phase voltages and phase currents of system

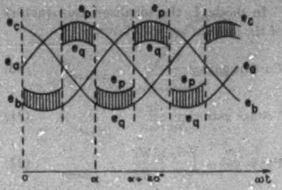


Fig. 5 (left). Voltages e_p and e_q of load terminals in relation to phase voltages e_a, e_b, and e_c at $\alpha=105$ degrees

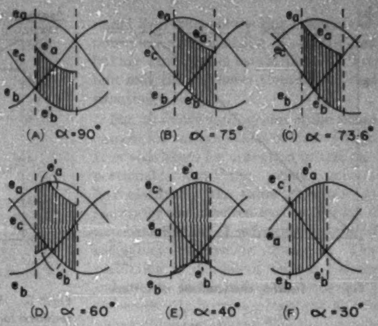


Fig. 6 (right). Voltages e_a' and e_b' relative to phase voltages over range $\alpha<\omega t<\alpha+60$ for various values of α; load voltage shaded

If the load resistance R_L in Fig. 1 is made equal to zero, then

$$e_p=e_q=e_{a'}=e_{b'}=e_{c'} \tag{5}$$

The phase-to-phase supply voltages must be taken up by the cores. But, from Fig. 3, the cores can absorb voltage only when the phase currents are $\pm N_oI_o/N_G$. Also, the sum of the phase currents must be zero.

$$i_a+i_b+i_c=0 \tag{6}$$

The only possible mode of operation which fulfils these conditions is two phases conducting currents of equal magnitude and opposite direction while the third has zero current. Each phase will conduct for a 120-degree period followed by a 60-degree zero-current period as shown in Fig. 4.

Measure angles from the reference angle when the phase voltage e_a is zero and increasing and designate the angle at which phase a switches to a conducting condition as α. This angle α can easily be shown to be 120 degrees for a short-circuited load since, for this condition, the average value of the phase (a) power ($e_a\ i_a$) must be zero.

Now relate the core-flux linkages to the applied voltages. At angle α, phase a begins conduction; the flux linkage $\lambda_{aa'}$ of phase a is zero. Phase b, having conducted negatively for the previous 60 degrees, will have a flux linkage which can be designated as $-\lambda_o$. In the interval $\alpha<\omega t<\alpha+60$, phase c will not conduct making $e_{c'}=e_c$. Thus the cores of phase a will have to absorb the line-to-line voltage e_{ac} over this interval and, in doing so, will acquire a flux linkage at $\omega t=\alpha+60$ of $\lambda_{aa'}=\lambda_o$.

If

$$e_a=E_m \sin \omega t \tag{7}$$

$$e_{ac}=\sqrt{3}\,E_m \sin (\omega t-30) \tag{8}$$

and

$$\lambda_o=\frac{1}{\omega}\int_{\alpha}^{\alpha+60} e_{ac}d\omega t$$
$$=\frac{1.5E_m}{\omega} \tag{9}$$

The supply voltage must be limited to a value which will make λ_o less than the value $2\lambda_m$ in Fig. 3 if the system is to operate at short circuit, since otherwise all cores would be saturated simultaneously. Thus, the peak phase voltage E_m is limited by the constraint:

$$E_m\leq 1.33\omega\lambda_m$$
$$E_m\leq 1.33N_GB_sA \tag{10}$$

From Figs. 4 and 1 it can be seen that the current in the short-circuited load has a constant, non-time-varying magnitude of

$$i_L=\frac{N_o}{N_G}I_o$$

Operation with Resistive Load

Suppose the system in Fig. 1 has a load resistance R_L and a control current I_o. Let phase a begin to conduct at angle α as shown in Fig. 4. During the interval $\alpha<\omega t<\alpha+60$ the return path is through phase b and $e_{cc'}=0$. In this interval, conduction occurs through a single loop $aa'pqb'b$ in Fig. 1 which can be described by the equation

$$e_{aa'}+R_Li_L-e_{bb'}=e_{ab}$$
$$=\sqrt{3}\,E_m \sin (\omega t+30) \tag{11}$$

The variation in the switching angle α as the load voltage changes may be derived by noting that at angle α, $\lambda_{aa'}=0$ while $\lambda_{bb'}$ is some negative value designated as $-\lambda_o$. At angle $\alpha+60$, $\lambda_{aa'}=0$ because of the phase symmetry, while $\lambda_{bb'}=0$. Thus, the flux linkages of both phases increase by λ_o during the period $\alpha<\omega t<\alpha+60$. Integrating both sides of equation 11 over this interval to obtain the flux linkages equivalent to the voltages gives

$$\frac{1}{\omega}\int_{\alpha}^{\alpha+60}(e_{aa'}+R_Li_L-e_{bb'})d\omega t$$
$$=\frac{1}{\omega}\int_{\alpha}^{\alpha+60} e_{ab}\ d\omega t$$

or

$$\lambda_o+\frac{R_Li_L}{\omega}\ \frac{\pi}{3}-\lambda_o=\sqrt{3}\,\frac{E_m}{\omega}\sin (\alpha+60)$$

Thus the load voltage e_{pq} can be expressed as a function of α.

$$e_{pq}=R_Li_L$$
$$=\frac{3\sqrt{3}}{\pi}E_m \sin (\alpha+60) \tag{12}$$

This load voltage can have a maximum average value of 1.65 at a switching angle α of 30 degrees. It will be shown later that the load voltage is non-time-varying up to a value of 1.63 E_m at a switching angle α of 40 degrees.

Now consider the load and core voltages in detail to determine the limitations on the constant current mode of operation. For the period $\alpha<\omega t<\alpha+60$, both cores of phase c are saturated, $i_c=0$ and $e_{c'}=e_c$. Conduction occurs through the rectifiers in branches $a'p$ and qb' only. If the rectifiers of phase c are not to conduct

$$e_p\geqslant e_{c'}\geqslant eq \tag{13}$$

In order to determine the voltages $e_{aa'}$ and $e_{bb'}$ across the core pairs in phases a and b, the voltages e_p and e_q must be known since $e_p=e_{a'}$ and $e_q=e_{b'}$. Within the limitation imposed by equation 13, the voltages e_p and e_q will be such as to make the voltages absorbed by the cores a minimum. Fig. 5 shows the voltages e_p and e_q in relation to the phase voltages for a switching angle of $\alpha=105$ degrees. Over the interval $\alpha<\omega t<\alpha+60$, $e_q=e_{c'}=e_c$ and

$$e_p=e_c+R_Li_L \tag{14}$$

The voltages $e_{aa'}$ and $e_{bb'}$ absorbed by the cores in phases a and b over this interval can now be described by:

$$e_{aa'}=e_{ap}$$
$$=e_{ac}-R_Li_L \tag{15}$$

$$e_{bb'}=e_{bq}=e_{bc} \tag{16}$$

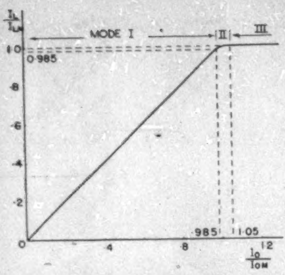


Fig. 7. Transfer characteristic of system

$$I_{Lm}=\frac{1.65E_m}{R_L};\qquad I_{om}=\frac{N_G}{N_o}I_{Lm}$$

During the interval of 60 degrees, the voltages $e_{aa'}$ and $e_{bb'}$ are absorbed by the unsaturated cores a_2 and b_1 respectively, cores a_1 and b_2 being in a saturated condition. The currents will be

$$i_a=-i_b=i_L=\frac{N_oI_o}{N_G}\tag{17}$$

Fig. 6 shows the load voltage in relation to the phase voltages for various values of the firing angle α all in the period $\alpha<\omega t<\alpha+60$ degrees. Note from Fig. 6(B) that, for $\alpha<90$ degrees, the voltage $e_{bb'}$ is initially negative. The flux linkage $\lambda_{bb'}$, which has a value of $-\lambda_o$ at angle α, becomes a larger negative value before increasing to zero at angle $\alpha+60$. Eventually, in Fig. 6(C), $e_p=e_a$ at angle α. From equation 12 this can be shown to occur at $\alpha=73.6$ degrees. Now, at angle α, $\lambda_{aa'}=0$ and i_a has a positive value. From Fig. 3, in the interval $\alpha<\omega t<\alpha+60$, $\lambda_{aa'}$ can only increase positively or remain at zero. It cannot become negative while carrying a positive current. Thus, for switching angles of $\alpha<73.6$ degrees, $e_{aa'}=0$ until the point in the interval when $e_{ac}=R_Li_L$ as shown in Fig. 6(D).

The load voltage e_{pq} remains constant in magnitude as the switching angle is decreased until the load voltage becomes equal to the voltage e_{ab} as shown in Fig. 6(E). Using equation 12, the switching angle α for this condition is found to be 40 degrees and the value of the load voltage is

$$e_{pq}=1.63E_m$$

If either the load resistance R_L or the control current I_o is increased further, the switching angle α will decrease from $\alpha=40$ degrees toward $\alpha=30$ degrees, see Figs. 6(E) and (F), and the load voltage e_{pq} will be equal to the voltage e_{ab} during part or all of the interval $\alpha<\omega t<\alpha+60$. At $\alpha=$ 30 degrees, $e_{pq}=e_{ab}$ throughout the interval of 60 degrees and the output voltage has a waveform identical with that of a 3-phase rectifier since all cores remain saturated.

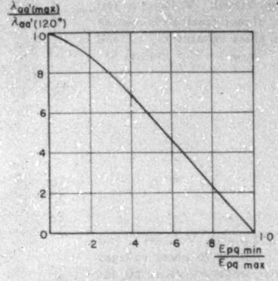


Fig. 8. Maximum flux linkages $\lambda_{aa'max}$ required for operation in limited load-voltage range from $e_{pq\,min}$ to $e_{pq\,max}$; $e_{pq\,max}=1.65\,E_m$; pu base for flux linkages is λ_o in equation 9

In this condition, e_{pq} varies between 1.5 E_m and 1.732 E_m and has an average value of 1.65 E_m. Thus, it is only in the upper 1.5% of the output voltage range between 1.63 E_m and 1.65 E_m that the output voltage and current cease to be non-time-varying.

SYSTEM TRANSFER FUNCTION

Over the output voltage range from zero to 1.63 E_m the load current I_L is constant and is related to the control current by

$$I_L=\frac{N_o}{N_G}I_o\tag{18}$$

To achieve the maximum average output voltage of 1.65 E_m the control current must be sufficient to allow a peak-load voltage of 1.732 E_m. Thus, the control current for maximum output is

$$I_o\geq\frac{1.732}{1.65}\frac{N_G}{N_o}I_L=1.05\frac{N_G}{N_o}I_L\tag{19}$$

where I_L is the average value of the output current.

The steady-state transfer function of the current source is shown in Fig. 7 on a per-unit (pu) basis. In mode I, the load current is directly proportional to the control current as stated in equation 18. In mode II, the load current increases by 1.5% with increased control current, and its ripple components increase rapidly. In mode III, all cores are continually saturated by the control current and the output is that of an uncontrolled 3-phase rectifier.

In mode I, the idealized gain factors of the system are as follows:

$$\text{Current gain}=\frac{N_o}{N_G}\tag{20}$$

$$\text{Voltage gain}=\frac{N_o}{N_G}\frac{R_L}{R_o}\tag{21}$$

$$\text{Power gain}=\left(\frac{N_o}{N_G}\right)^2\frac{R_L}{R_o}\tag{22}$$

where R_o is the total resistance of the control circuit.

OPERATION OVER LIMITED OUTPUT VOLTAGE RANGE

When the load consists of a short circuit, the cores must absorb all of the applied voltage. For this condition the supply voltage is limited in its maximum value as given in equation 10. At the maximum value of output voltage, no voltage is absorbed by the cores. In some situations a system may be desired which will control the output current into a load whose voltage will vary only between its maximum value and some fraction of the maximum value. Determine the relationship between the maximum supply voltage E_m and the maximum flux linkage λ_m of a single core for this type of operation. To do this, first determine the maximum value of the flux linkage $\lambda_{aa'}$ of phase a as a function of α. At $\omega t=\alpha$, $\lambda_{aa'}=0$. For any value of $\omega t>\alpha$ the flux linkage $\lambda_{aa'}$ is given by:

$$\lambda_{aa'}=\frac{1}{\omega}\int_\alpha^{\omega t}e_{aa'}\,d\omega t\tag{23}$$

and will increase as long as $e_{aa'}$ is posi-

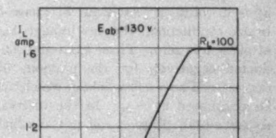


Fig. 9. Experimental curves of load current/control current for various values of load resistance R_L

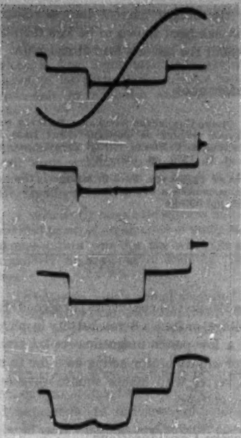


Fig. 10. Phase current (i_a) waveforms for load-current values of 0.41, 0.66, 0.77, 1.0 pu of maximum load current; voltage e_{ab} shown as reference

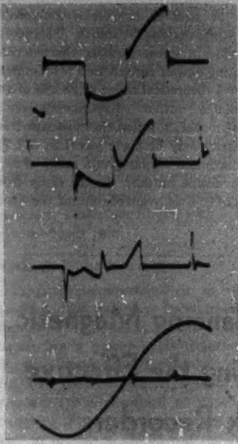


Fig. 11. Phase voltage ($e_{aa'}$) for same conditions as Fig. 10

tive. For switching angles in the range $120>\alpha>90$ as shown in Figs. 5 and 6, $e_{aa'}$ is positive only over the region $\alpha<\omega t<\alpha+60$. Substituting into equation 23 the value of $e_{aa'}$ from equation 15 and integrating.

$$\lambda_{aa'\ \max}=\frac{\sqrt{3}E_m}{\omega}\sin(\alpha-60)\text{ for } 120>\alpha>90 \quad (24)$$

For $\alpha<90$ note from Fig. 6 that $e_{aa'}$ is positive for a time after $(\alpha+60)$. For example, in Figs. 6(B) and (C), $e_{aa'}$ is equal to e_{ab} over the interval $\alpha+60<\omega t<150$. Evaluation of the maximum value of flux linkage $\lambda_{aa'}$ including this additional component gives

$$\lambda_{aa'\ \max}=\frac{\sqrt{3}E_m}{\omega}[1-\sin(\alpha+60)]\text{ for } 90>\alpha>73.6 \quad (25)$$

By a similar integration, the value of $\lambda_{aa'\max}$ may be found for the range $73.6>\alpha>30$ degrees.

A simultaneous solution of equations 24 or 25 and 12 gives the maximum flux linkage in phase a as a function of the load voltage e_{pq}. This is shown in Fig. 8, using normalized co-ordinates. If, for example, the application requires the load voltage to be controllable only from a maximum value of 1.65 E_m down to 0.5 of this value, the maximum flux linkage required in the cores would be only 0.577 of the value required if operation down to zero output voltage were required. Equations 9 and 10 give the required flux linkages per core for operation at zero output voltage. For operation over a limited voltage range, the pu reduction in required flux linkage per core is then given directly by the ordinate of Fig. 8.

Experimental Results

Experiments were performed on the system shown in Fig. 1, using six toroidal cores of Deltamax* material, each having a saturation voltage of 65 volts and a turns-ratio $N_o/N_G=2.25$. Fig. 9 shows a set of experimental curves for output current I_L plotted against control current I_o for various values of load resistance R_L. These curves are seen to be of the same general shape as the ideal normalized curve of Fig. 7. The load current is linearly proportional to the control current over most of the proportional (mode I) range. The departure from a linear relationship for low values of load current is caused by the magnetizing current, about 0.03 ampere, of the unsaturated cores. With careful design it is considered possible to reduce this magnetizing current to less than 0.5% of the rated load current of the core windings. The departure from linearity in the high-output-current region of mode I is caused by leakage inductance which tends to round the corners of the flux-linkage/current characteristic of Fig. 3. If the flux linkage excursion is small, the load current may then be lower than in the ideal situation.

Fig. 10 shows oscillograms of the phase current for various values of load voltage. It is seen that, for operation in the proportional range, the phase current has essentially the same waveform as shown for the idealized system in Fig. 4. After the instant of switching, a damped oscillatory current is observed. This oscillation is dependent on the leakage inductance and winding capacitance of the cores. It is most prominent at low values of output voltage where the damping is a minimum.

The load current has the same waveform as a typical 120-degree conduction section of the phase current shown in Fig. 10. Over the range from 5% to 95% of maximum output voltage it was found that the ripple on the output current was less than 1% with a resistive load. This ripple may arise from several sources. One easily controlled source of ripple is an unbalance in the turns-ratios of the cores which causes a cyclical variation in the magnitude of the output current. It is therefore important that the numbers of turns in all windings be exact. Second, the magnetizing currents of the cores contribute a component of ripple which may be minimized by the use of a core material with high unsaturated permeability and a core with a minimum mean length of magnetic path. Third, a high-frequency ripple component arises from the oscillation following switching as shown in Fig. 10. This component can easily be reduced by the introduction of a small inductor in series with the load, or a small capacitor in parallel with the load. If very low ripple is required a conventional filter may be inserted in the load circuit, but, for most applications, the unfiltered output expected is adequate.

Fig. 11 shows the voltage $e_{aa'}$ across the two cores in phase a for various values of output voltage. These are seen to be similar to the waveforms which are derivable from Fig. 6 with the exception of the high-frequency oscillation following the instant of switching, particularly at low values of output voltage. These oscillatory voltage peaks add to the inverse voltage across the rectifiers and they must therefore be controlled or allowed for in the selection of the peak inverse rating of the rectifiers.

Since the problems of balancing and of reduction of the magnetizing current in relation to the rated current are more difficult in small cores, Deltamax material was used in the model which was tested. However, it is contemplated that larger units can be constructed satisfactorily using grain-oriented transformer-grade materials.

---

* Trade name, Arnold Engineering Company, Marengo, Ill.

Conclusions

A system using a 3-phase supply, saturable magnetic cores, and rectifiers can be used to provide a controllable non-time-varying d-c source. This source provides a linear transfer characteristic relating output current to control current over a theoretical range of 98.5% of the available output. Experiments on a small model show that the departure from linearity may be made less than ±1.5% of maximum output over a range of 95% of available output. In the range from 5% to 95% of available output the ripple on the output current into a resistive load has been shown to be less than 1% without the use of a load filter.

References

1. SERIES-CONNECTED SATURABLE REACTOR WITH CONTROL SOURCE OF COMPARATIVELY HIGH IMPEDANCE, H. F. Storm. *AIEE Transactions*, vol. 69, pt. II, 1950, pp. 1299–1309.

2. THE TRIDUCTOR, Paul P. Biringer. *Ibid.*, pt. I (*Communication and Electronics*), vol. 75, Nov. 1956, pp. 590–94.

Single-Ended Self-Balancing Magnetic Amplifiers for Changing the Effective Impedance of Ink Recorders

WILLIAM A. GEYGER
FELLOW AIEE

ANNOUNCEMENT was made in 1952 of a magnetic amplifier of the self-balancing potentiometer type[1] combining the very stable operating conditions of d-c instrument transformers[2,3] with the unique features of self-balancing potentiometer devices and negative-feedback amplifiers. Thus the door was opened to widespread application of such "self-balancing" magnetic amplifiers[4] in fields which are far beyond the capabilities of the conventional magnetic-amplifier circuits. Actually, since then, this subject has engaged the attention of many research and development groups with regard to design[5,6] and analytical investigations.[7–10] As a result of the development of various types of self-balancing magnetic-amplifier push-pull circuits utilizing either differential external feedback[11,12] or self-saturation,[13,14] the field of application is continually being broadened.

Recently, in conjunction with development of a miniaturized d-c instrument transformer,[15] further research work has been conducted toward refinements and improvements on single-ended self-balancing magnetic amplifiers operating a standard permanent-magnet moving-coil ink recorder. The aim of this activity consisted essentially in providing a simple and reliable impedance-changing device for operating the Esterline-Angus Company, Inc., 1-ma (milliampere) 1,500-ohm recorder as a high-impedance voltmeter or as a low-impedance milliammeter, respectively. Such a device makes it possible to use the standard recorder for functions where the power of 1.5 mw (milliwatts) absorbed from the circuit under test and the associated voltage drop of 1.5 volts have hitherto proved prohibitive.

This paper presents detailed design information and performance characteristics of a 60-cycle-operated prototype form of a single-ended self-balancing magnetic amplifier operating either under voltage-balance or current-balance conditions with infinite internal gain. The amplifier acts as an impedance-changing device and permits operation of the standard recorder as a high-impedance voltmeter (input current 0–2 μa [microamperes] with the voltage ranges 0–0.25 volts to 0–10 volts) or as a low-impedance milliammeter (voltage drop 0–7 mv [millivolts] with the current range 0–1 ma) without impairing the accuracy of measurements and without affecting the dynamic performance of the recorder. The amplifier may be supplied either from the 120-volt 60-cps (cycles per second) line voltage, or from a 6-volt battery by means of a low-power magnetic-coupled transistor multivibrator acting as a d-c to a-c converter with 60-cps square-wave output voltage.

Design Considerations

The specific problem consisted in operating the Esterline-Angus 1-ma 1,500-ohm recorder from a self-balancing magnetic amplifier acting as an impedance-changing device in such a way that the recorder can be employed as a high-impedance voltmeter or as a low-impedance microammeter, respectively. When measuring voltages in the ranges 0–0.25 volt to 0–10 volts, the effective input resistance of the amplifier should be so high that not more than about 3 to 5 μa will be drawn from the d-c signal source having a comparatively high internal resistance of about 5 to 20 kilohms. With regard to measurement of currents in the range 0–1,000 μa, the effective input resistance of the amplifier should be so low that the voltage drop introduced into the circuit under test does not exceed about 5 to 10 mv.

Because the maximum load-current value corresponding to full-scale deflection of the recorder is 1,000 μa, the magnetizing-current component of the saturable-reactor elements should be of the order of 1% of this value, that is, about 10 μa, or less. In order to reduce the magnetizing current to the minimum, the following aspects, similar to those concerning the design of a miniaturized d-c instrument transformer,[15] are considered to be of prime importance:

1. Although the rectangularity of the dynamic hysteresis loop of the magnetic core materials with very low coercive force (H_c), e.g., supermalloy or HyMu *80*, is less pronounced than that of the rectangular hysteresis-loop grain-oriented nickel-iron alloys, the low-coercive-force materials are definitely preferred in this special application in order to minimize the magnetizing-

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted May 2, 1960; made available for printing September 21, 1960.

WILLIAM A. GEYGER is with the U. S. Naval Ordnance Laboratory, Silver Spring, Md.

The author expresses his thanks to Horace H. Helms, Edward T. Hooper, and Michael Pasnak for their helpful co-operation in conducting the experimental investigations leading to this paper.

current requirements of the self-balancing magnetic amplifier.

2. Although the power-supply frequency, 60 cps, is low, 2-mil tape cores are considered to be particularly suitable in this application to minimize eddy-current effects.

3. The 60-cycle-operated saturable-reactor elements must be designed to have a very small cross-sectional area and a correspondingly high number of turns of their gate windings.

4. A sufficiently large core window must be provided to avoid application of undesirable extremely thin wire for the windings and excessively high values of winding resistance.

5. In order to reduce actual magnitude of the magnetizing-current component, it is desirable to provide an a-c bias circuit fed from the power-supply voltage.

In accordance with common practice in the design of magnetic amplifiers, it is advantageous to use two individually wound gate windings and several control and feedback windings which embrace both magnetic circuits.[16] With this arrangement, no net voltage of fundamental frequency can appear across the common-wound control and feedback windings. To minimize the magnetizing-current requirements, Mo-permalloy 2-mil tape cores are used in the prototype design. Fig. 1 presents the 60-cps hysteresis loop of these cores, which has been automatically ink-recorded by means of a simple dynamic **B-H** loop tracer with silicon-diode chopper circuits (Ferrotracer),[17] and a chopper-type x-y plotter (Moseley Autograf, F. L. Moseley Company). The magnetic-core test circuit itself has been designed so that the core sample is tested under sinusoidal-flux excitation conditions.

Voltage-Type Self-Balancing Magnetic Amplifier

In accordance with the fundamental principle of the voltage-type self-balancing magnetic amplifier, the d-c voltage to be measured is balanced in the control-circuit loop by an opposing compensating voltage, which is derived with negative "electric" feedback from the output circuit of the amplifier; and, additionally, positive "magnetic" feedback is applied to obtain infinite internal gain or "critical regeneration."[1,4] Various forms of full-wave magnetic-amplifier circuits, single-ended or push-pull, utilizing either external feedback or self-saturation, may be employed for this purpose. In any case, the amount of the applied positive magnetic feedback will be adjusted so that the average value of the control-winding current is substantially zero. The very large amount of electric negative feedback used makes the external gain of the amplifier essentially independent of all factors except the accuracy of the resistors in the negative-feedback network. Thus, over certain ranges, large variations in magnitude, frequency and waveshape of power-supply voltage, load resistance, and control-circuit resistance have no appreciable effect upon the transconductance or voltage ratio of the amplifier operating under extremely stable conditions with a high over-all power gain.

VOLTAGE-BALANCE CIRCUIT

From the many possible circuit configurations of full-wave-type magnetic amplifiers the self-saturating circuit of Fig. 2, containing a power-supply transformer T_P with center-tapped secondary winding, has been selected for the following reasons: (1) With regard to impedance matching of circuit components it is necessary to provide a 120/30-volt step-down transformer. (2) Because the output circuit containing the ink recorder (load R_L) is here resistance-coupled with the control-circuit loop by means of the compensating resistance R_K, it is desirable to isolate the output circuit from the power supply ($E_P = 120$ volts, $f_P = 60$ cps). (3) Application of a center-tapped secondary winding of T_P makes it possible, in accordance with common practice, to operate with two half-wave rectifier units (silicon diodes, type *1N646*) only, when employing an external-feedback circuit or a bridge-type self-saturating circuit, four such rectifier units must be provided.

The circuit of Fig. 2 employs two equally rated saturable-reactor elements with gate windings N_A, a-c bias windings N_B, d-c control windings N_C, and d-c feedback windings, N_D. The d-c signal voltage E_S, i.e., the voltage to be measured, is balanced by the compensating voltage $E_K = I_L R_K$. The average value of control-winding current is, if I_C is very small in comparison with the unidirectional load current I_L,

$$I_C = \frac{E_S - E_K}{R_C + R_M} = \frac{E_S}{R_C + R_K} - \frac{E_K}{R_C + R_K} = I_S - I_K \tag{1}$$

where E_S is the electromotive force of the d-c signal source and R_C is the total resistance of the control circuit between the terminals of negative-feedback resistor R_K (including the sum of the resistances of the control-winding units N_C and the internal resistance of the signal source). The term $I_S = E_S/(R_C + R_K)$ is the signal-current component, and $I_K = E_K/(R_C + R_K)$ is the compensating-current component.

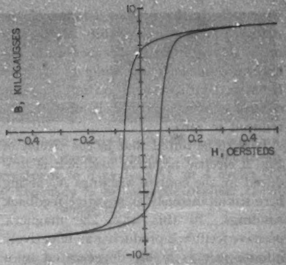


Fig. 1. B-H loop, 60 cps, of Mo-permalloy 2-mil tape cores used in prototype design of single-ended self-balancing magnetic amplifiers with circuit of Fig. 2 or Fig. 6

It has been shown[1,18] that, with critical regeneration operating conditions,

$$\frac{I_L}{E_S} = \frac{1}{R_K} \tag{2}$$

where $1/R_K$ represents the transconductance of the circuit. In this case

$$I_C = I_S - I_K = 0 \tag{3}$$

Critical regeneration (infinite internal gain of the amplifier) is obtained by magnetic positive feedback produced by self-saturation (gate windings N_A with series-connected half-wave rectifiers) and by the small additional positive-feedback effect of external-feedback windings N_D carrying the positive-feedback current I_D. Actual current ratio I_D/I_L can be conveniently adjusted by a current-divider potentiometer resistor (Ayrton-Mather shunt) R_D:

$$\frac{I_D}{I_L} = \frac{R_E}{R_D} \tag{4}$$

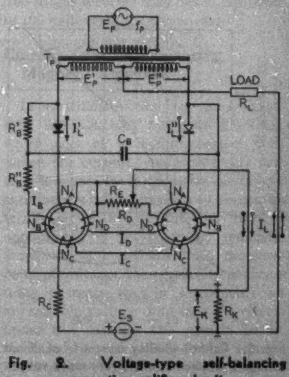


Fig. 2. Voltage-type self-balancing magnetic-amplifier circuit

where R_D is the total resistance of the regeneration-control circuit N_DR_D. In order to obtain critical-regeneration operating conditions, the slider of R_D is to be moved into such a position that the control-winding current I_C is substantially reduced to zero.

It is to be noted that in the circuit of Fig. 2 electric negative feedback is produced by the control windings N_C acting here simultaneously as negative-feedback windings. In this way, the magnetic positive feedback producing an effectively infinite gain is highly degenerated by a very large amount of negative electric (galvanic) resistance coupled into the control circuit loop. As a result, the effective input resistance of the amplifier is about 100 to 1,000 times higher than the actual input-circuit resistance. In this form, the self-balancing magnetic amplifier measures voltages in such a way that no appreciable current (I_C) is drawn from the signal source. Balance condition is fulfilled, if the average value of the compensating voltage drop $E_K=I_LR_K$ is equal and opposite to the actual value of the voltage E_S to be measured. Thus, in this case, the amplifier represents a voltmeter having a very high resistance (E_S/I_C).

The effective or total feedback factor of the circuit of Fig. 2 involves several components. This factor is a measure for the stability of operation. In accordance with Storm's concept,[19] the magnitude of the intrinsic negative feedback of the ordinary saturable reactor can be expressed by the intrinsic-feedback factor $h_i=-1$. This negative feedback is here completely compensated or nullified by the positive magnetic feedback (produced by self-saturation and by the external-feedback windings N_D) which can be represented by the positive extrinsic-feedback factor $h_{e,\,m}=+1$. The negative electric feedback which is produced by the control windings N_C (acting here as negative-feedback windings) can be expressed by the negative extrinsic-feedback factor $h_{e,e}$. Thus the effective or total feedback factor h of the circuit of Fig. 2 is defined as follows:[18]

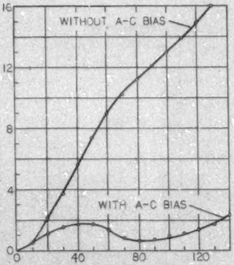


Fig. 3. Control-winding current I_C of circuit of Fig. 2 as function of power-supply voltage E_P

$$h_i=-1,\ h_{e,m}=+1$$

and

$$h_{e,e}=-2\frac{N_C}{N_A}\frac{R_K}{R_C+R_K} \tag{5}$$

Therefore

$$h=h_i+h_{e,m}+h_{e,e}=-2\frac{N_C}{N_A}\frac{R_K}{R_C+R_K} \tag{6}$$

Furthermore, the remaining amount of applied positive feedback may be characterized by the resultant extrinsic-feedback factor:

$$h_e=h_{e,m}+h_{e,e}=1-2\frac{N_C}{N_A}\frac{R_K}{R_C+R_K} \tag{7}$$

When expressing the remaining amount of applied positive feedback in per cent,[20] the following equation may be used:

$$F_{eff}=+100\left(1-2\frac{N_C}{N_A}\frac{R_K}{R_C+R_K}\right)\text{ per cent} \tag{8}$$

COMPENSATION OF MAGNETIZING-CURRENT COMPONENT WITH A-C BIAS CIRCUIT

With no-signal conditions ($E_S=0$), the very small quiescent-current value of I_L causes a voltage drop $E_K=I_LR_K$ across R_K, and a very small negative control-winding current $I_C=I_LR_K/(R_C+R_K)$ flows through the control windings N_C. Note that the voltage-balance circuit assumes a conduction path through the voltage to be measured so that this negative control-winding current producing negative feedback can flow through the control-circuit loop.

In order to reduce the quiescent-current value of I_L to the minimum, an a-c bias circuit, as shown in Fig. 2, may be used. Since the magnetomotive force required by a nickel-iron-alloy core over intervals of increasing or decreasing fluxes is practically constant, a large part of it may be provided by an additional a-c winding fed from a synchronous bias current I_B which is derived from the power-supply voltage E_P. However, with low-coercive-force core materials (supermalloy, HyMu 80), it is possible to obtain the desired reduction of the magnetizing-current component only if the phase relationship between I_B and E_P can be adjusted to the proper value by means of a correspondingly rated reactor or capacitor in the a-c bias circuit. In this case, adjustment of the respective components of the a-c bias circuit must be made so that the fundamental wave of the magnetizing current and that of the bias current I_B are in phase, in order to bring the magnetizing-current component of I_L to the minimum (see pp. 77–81 of reference 16). In the circuit of Fig. 2, bias resistors R_B', R_B'' and shunt capacitor C_B must be properly rated to obtain optimum value of phase displacement between the fundamental waves of bias current I_B and secondary voltage $(E_P'+E_P'')$ of the power-supply transformer T_P.

PERFORMANCE CHARACTERISTICS OF PROTOTYPE DESIGN

In accordance with the dimensions of components of the miniaturized d-c instrument transformer described previously,[15] the following core and winding data have been used in a prototype design based upon the circuit of Fig. 2:

Mo-permalloy 2-mil tape cores, tape width = 0.125 inch
Inside diameter = 1.250 inch, outside diameter = 1.500 inch
Gate windings (individually-wound), N_A: 20,000 turns of no. *40* Awg (American wire gage) wire, 3,000 ohms each
Bias windings (individually wound), N_B: 1,000 turns of no. *40* Awg wire, 190 ohms each
Control windings (common wound), N_C: 10,000 turns of no. *40* Awg wire, 3,400 ohms
Feedback windings (common wound), N_D: 1,000 turns of no. *40* Awg wire, 360 ohms
Silicon diodes: type *1N646* (or equivalent)

Actual voltage range of the amplifier will be changed simply by varying the negative-feedback resistor R_K. When employing the Esterline-Angus recorder (I_L=0–1 ma), the voltage ranges 0–1 volt, 0–5 volts, and 0–10 volts can be obtained with R_K=1,000 ohms, 5,000 ohms, and 10,000 ohms, respectively. Smaller voltage ranges, e.g., 0–0.25 volt (having correspondingly reduced accuracy of measurement) are possible with sufficient stability of operation. Higher voltage ranges, up to about 10,000 volts, may be provided by means of a voltage divider (1,000 ohms/volt) with modern types of high-resistance components.

At first, the magnetizing-current component of load current I_L was minimized by systematically varying and mutually readjusting the components R_B' R_B'', and C_B of the a-c bias circuit. In this procedure, the average value of control-winding current I_C has been measured by means of a moving-coil lightspot galvanometer

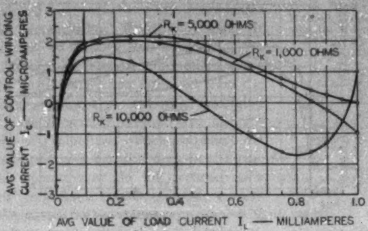


Fig. 4 (left). Control-winding current I_C of circuit of Fig. 2 as function of load current I_L with negative-feedback resistance R_K as parameter

Fig. 5. Waveform of load current I_L with half-cycle-response operating conditions; sweep frequency of cathode-ray oscilloscope corresponds to 8 cycles (A), or 4 cycles (B) of power-supply frequency f_P of magnetic-amplifier circuit under test

(−10 to 0 to +10 μa, 15 ohms) which is series-connected with N_C. Fig. 3 shows, with no-signal conditions (E_S=0), actual average value of I_C as a function of the power-supply voltage $E_P=2(E_P'+E_P'')$, without a-c bias and with properly adjusted a-c bias. In both cases, the negative-feedback resistor is R_K=10,000 ohms corresponding to the largest voltage range, E_S=0–10 volts. Evidently, with E_P=120 volts, and optimum adjustment of the bias circuit (R_B'=160,000 ohms, R_B''=30,000 ohms, and C_B=0.05 μf [microfarads]), the negative control-winding current I_C (with E_S=0) has been reduced to about 1.4 μa, that is, to about 0.1 of its original value without a-c bias.

With regard to measurement of voltages, Fig. 2, the purpose of the self-balancing magnetic amplifier is to measure the direct voltage, E_S, applied to the input terminals in terms of the average value of load current I_L flowing through the recorder. Therefore, by far the most important test to be made on such an amplifier is to find how correctly the load current I_L is an exact measure of E_S. Since the average value of control-winding current I_C is an "error current" indicating actual deviation of the average value of load current I_L from its ideal value, E_S/R_K (equation 2), the percentage transconductance error, i.e., the deviation of the actual value of transconductance I_L/E_S from its ideal value, $1/R_K$, is

$$\Delta(I_L/E_S)=100\left(\frac{I_C}{I_L}\right)\text{ per cent} \qquad (9)$$

This error $\Delta(I_L/E_S)$ is positive when, with a given signal voltage E_S, the actual load current I_L is in excess of its ideal value, i.e., when the recorder reads too high.

Fig. 4 shows the average value of control-winding current I_C as a function of average value of load current I_L with negative-feedback resistance R_K as a parameter. In all cases, I_C varies within the limits of 0 to ±2 μa. The actual effective input resistance of the amplifier, as seen from the d-c signal source (E_S), is defined by the ratio E_S/I_C. Thus, E_S/I_C will have its highest value with R_K=10,000 ohms (E_S=0–10 volts) and will decrease with smaller value of R_K. Evidently, input-resistance values of the order of 0.10 to 10 megohms will be obtained in this way.

Further measurements on the prototype design using the circuit of Fig. 2 have shown that the influence of large variations of the internal resistance of the signal source within the limits of about 50 to 50,000 ohms upon the effective input resistance, E_S/I_C, is quite small, when applying the larger voltage ranges, e.g., 0–1 volt to 0–10 volts. However, it is to be noted that the dynamic performance (speed of response) of the circuit of Fig. 2 will be affected by increasing internal resistance of the signal source, i.e., by increasing total resistance R_C of the control circuit.[10]

The dynamic performance of the prototype design with the circuit of Fig. 2 has been studied by means of an improved form of transient analyzer[15] which permits direct visual observation of a stationary pattern on a cathode-ray oscilloscope. In this transient analyzer, a unidirectional square-wave signal voltage (E_S) is derived from a magnetic-coupled multivibrator which is synchronized with the power-supply frequency (f_P) of the magnetic-amplifier circuit under test. A unique feature of this arrangement resides in the fact that in contrast to conventional forms of switching transistor d-c to a-c converters,[21–24] the oscillation frequency of the multivibrator can be varied by means of a standard low-power continuously-variable transformer, within wide limits, without affecting the magnitude of the output voltage and without impairing the output waveform.[25–26] When operating this transient analyzer with 60 cps, the oscillation frequency of the multivibrator is at first adjusted to 30 cps by slowly varying the position of the slider of the variable transformer. Then, in order to obtain a stationary pattern on the screen of the oscilloscope, an auxiliary 60-cps unidirectional square-wave voltage (derived from a ferroresonant circuit with half-cycle splitting network) is injected into the base circuit of one of the two switching transistors. In this case, the multivibrator provides a synchronized 30-cps unidirectional square-wave signal voltage (E_S) at the input terminals of the magnetic-amplifier circuit under test. By mutually readjusting the slider of the variable transformer and that of a potentiometer resistor in the half-cycle splitting network of the ferroresonant circuit,[15] a stationary pattern can be displayed and photographically reproduced in the usual way.

The oscillograms of Fig. 5 illustrate the dynamic performance of the prototype design with the circuit of Fig. 2 in the special case that the negative-feedback resistance R_K has such a value (10,000 ohms corresponding to the voltage range E_S=0–10 volts) that half-cycle response is obtained. The sweep frequency of the oscilloscope has been synchronized so that 8 cycles, Fig. 5(A), or only 4 cycles, Fig. 5(B), of the output current I_L will appear on the screen. It is to be noted, however, that half-cycle-response operating conditions are generally not required in the present application and that the response time of the amplifier operating the Esterline-Angus recorder may be of the order of 100 milliseconds, or even more.

Current-Type Self-Balancing Magnetic Amplifier

In contradistinction to the voltage-type self-balancing magnetic amplifier, which acts as a high-impedance voltmeter, the current-type counterpart acts as a low-impedance ammeter having an extremely small voltage drop across its terminals. Although the majority of applications of self-balancing magnetic amplifiers is concerned with measurement of voltages, the possibility of measuring currents without introducing an appreciable voltage drop into the circuit under test carrying the current to be measured is considered to be equally important. Application of the current-type self-balancing magnetic amplifier

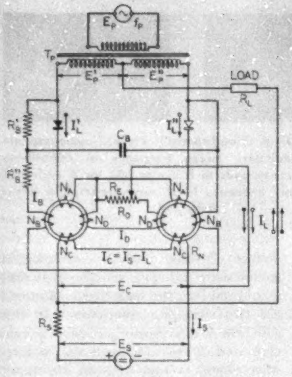
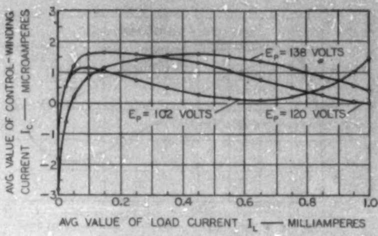


Fig. 6 (left). Current-type self-balancing magnetic-amplifier circuit

Fig. 7 (right). Control-winding current I_C of circuit of Fig. 6 as function of load current I_L with power-supply voltage E_P as parameter



in conjunction with the standard 1-ma 1,500-ohm ink recorder makes it possible to use this recorder for functions where the voltage drop of 1.5 volts has hitherto proved prohibitive.

CURRENT-BALANCE CIRCUIT

In the circuit of Fig. 6, the signal current I_S, i.e., the direct current to be measured, is balanced by the opposing load current I_L, and the control-winding current $I_C=I_S-I_L$ is substantially reduced to zero as a result of critical regeneration (infinite internal gain provided by self-saturation and the external-feedback circuit N_DR_D producing additional magnetic positive feedback). This arrangement operating under current-balance conditions of the control circuit is characterized by the fact that the effective input resistance of the magnetic amplifier is greatly reduced, and that the actual voltage drop, $E_C=I_CR_N=(I_S-I_L)R_N$, across the control windings N_C (resistance R_N) is substantially reduced to zero. Therefore, no undesirable excessive voltage drop will be introduced into the circuit under test carrying the direct current I_S to be measured.

The effective feedback factor of the circuit of Fig. 6 is given by:[18]

$$h_t=-1,\ h_{e,m}=+1$$

and

$$h_{e,e}=-2\,\frac{N_C}{N_A} \tag{10}$$

Therefore

$$h=h_t+h_{e,m}+h_{e,e}=h_{e,e}=-2\,\frac{N_C}{N_A} \tag{11}$$

In accordance with the prototype design, $N_C=N_A/2=10{,}000$, $h_{e,e}=-1$, and $F_{eff}=0$, i.e., the circuit has maximum stability of operation. It is to be noted that the total amount of load current I_L is superposed upon the signal current I_S. The unique feature of this circuit is the fact that the actual voltage drop, E_C, across the control windings N_C is reduced to an extremely small value which is about 1,000 times smaller than the corresponding voltage drop across the control windings of the ordinary saturable reactor of equivalent design.[18] However, the stability of operation, as characterized by the effective feedback factor $h=h_t=h_{e,e}=-1$, is the same in both cases.

PERFORMANCE CHARACTERISTICS OF PROTOTYPE DESIGN

The core and winding data of the saturable-reactor construction with the individually wound windings N_A and N_B and the common-wound windings N_C and N_D are the same as those of the voltage-type arrangement; and the same dimensions of components R_B' and R_B'' and C_B have been used in the a-c bias circuit to reduce the magnetizing-current component of load current I_L to the minimum.

With regard to measurement of currents, Fig. 6, the purpose of the self-balancing magnetic amplifier is to measure the current, I_S, applied to the terminals of control winding N_C in terms of the average value of load current I_L flowing through the recorder. In this case, therefore, by far the most important test to be made on the current-type amplifier is to find how correctly the load current I_L is an exact measure of I_S. Since the average value of control-winding current, $I_C=I_S-I_L$, is an "error current" indicating actual deviation of the average value of load current I_L from its ideal value, I_S, the percentage current-ratio error, i.e., the deviation of the actual value of current ratio from its value, $I_L/I_S=1$, is:

$$\Delta(I_L/I_S)=100\left(\frac{I_C}{I_L}\right)\text{ per cent} \tag{12}$$

This error Δ (I_L/I_S) is positive when with a given signal current I_S, the actual load current I_L is in excess of its ideal value, i.e., when the recorder reads too high.

The average value of control-winding current $I_C=I_S-I_L$ has been measured by means of a moving-coil lightspot galvanometer (-1 to 0 to $+10\,\mu$a, 15 ohms) which is series-connected with N_C (copper resistance $R_N=3{,}400$ ohms). Fig. 7 shows I_C as a function of I_L with power-supply voltage $E_P=2(E_P'+E_P'')$ as a parameter. Since I_C varies within the limits of 0 to $\pm2\ \mu$a, the voltage drop across N_C does not exceed 7 mv. Thus, actual voltage drop across this current-measuring impedance-changing device has been reduced to about 1/200 of its original value (1.5 volts representing the voltage drop across the recorder).

Fig. 7 makes it evident that the effect of the applied variations of ±15% of the 60-cps line voltage E_P is negligible. This is also true if the amplifier is supplied from a 6-volt battery by means of a low-power magnetic-coupled transistor multivibrator acting as a d-c to a-c converter with 60-cps square-wave output voltage.

At this point, it is interesting to note that the current-type self-balancing magnetic amplifier is also suitable for performing summation of several direct currents by means of a single control winding N_C. Because the voltage drop, E_C, across N_C is extremely small, there will be no undesirable interference between the systems carrying the individual signal currents to be added.

The dynamic performance of the prototype design with the current-type circuit of Fig. 6 has been studied by means of the transient analyzer[18] with a special form of magnetic-coupled multivibrator.[25,26] The oscillograms obtained in this case are identical with those of Fig. 5 corresponding to half-cycle-response operating conditions of the voltage-type circuit of Fig. 2 with $R_K=10{,}000$ ohms.

Conclusions

Sixty-cycle-operated single-ended self-balancing magnetic amplifiers using low-coercive-force nickel-iron-alloy core material (supermalloy, HyMu *80*) make it possible to employ the standard 1-ma 1,500-ohm ink recorder for functions where the power of 1.5 mw absorbed from the circuit under test and the associated voltage drop of 1.5 volts have hitherto proved prohibitive. Such amplifiers act as impedance-changing devices and permit operation of the ink recorder as a high-impedance voltmeter (input current 0–2 μa with the voltage ranges 0–0.25 volt to 0–10 volts) or as a low-impedance milliammeter (voltage drop 0–7 mv with the current range 0–1 ma) without impairing the accuracy of the recorder. The amplifiers may be supplied either from the 120-volt 60-cps line voltage, or from a 6-volt battery by means of a low-power magnetic-coupled transistor multivibrator acting as a d-c to a-c converter with 60-cps square-wave output voltage.

The dynamic performance of the prototype design described in this paper (half-cycle-response operating conditions) makes it evident that these special forms of magnetic amplifiers may also be useful for various other problems in the realm of d-c instrumentation.

References

1. Magnetic Amplifiers of the Self-Balancing Potentiometer Type, W. A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 71, Jan. 1952, pp. 383–95.

2. A Simple D-C Instrument Transformer Having Real Current-Transformer Properties (in German), W. Krämer. *Elektrotechnische Zeitschrift*, Wuppertal-Elberfeld, Germany, vol. 58, 1937, pp. 1309–13.

3. A New D-C Instrument Transformer for the Measurement of High D-C Voltages (in German), W. Krämer. *Ibid.*, vol. 59, 1938, pp. 1295–98.

4. Self-Balancing Magnetic Amplifier, W. A. Geyger. *U. S. Patent no. 2,700,130*, Jan. 18, 1955.

5. An Operational-Type Magnetic Amplifier for Air-Borne Servo-Control Systems, R. M. Hubbard. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 73, Nov. 1954, pp. 425–30.

6. An Analogue Computer Technique Using Magnetic Amplifiers, B. E. Davis, I. H. Swift. *Ibid.*, vol. 73, 1954 (Jan. 1955 section), pp. 635–40.

7. Fast Response with Magnetic Amplifiers, D. G. Scorgie. *Ibid.*, vol. 72, 1953 (Jan. 1954 section), pp. 741–49.

8. The Operation of the Self-Balancing Magnetic Amplifier, Albert D. Krall, Edward T. Hooper. *Ibid.*, vol. 75, Mar. 1956, pp. 79–84.

9. The Operation of the Current-Type Self-Balancing Magnetic Amplifier, Albert D. Krall. *Ibid.*, vol. 76, Sept. 1957, pp. 380–84.

10. On the Feedback in Magnetic Amplifiers; II—Combined Magnetic and Electric Feedbacks, L. A. Finzi, J. J. Suozzi. *Ibid.*, vol. 78, May 1959, pp. 136–41.

11. Self-Balancing Magnetic Amplifiers of the Differential-Feedback Type, William A. Geyger. *Ibid.*, vol. 76, Mar. 1957, pp. 39–46.

12. Sixty-Cycle Low-Level Self-Balancing Magnetic Amplifier of the Differential-Feedback Type, William A. Geyger. "Conference on Magnetic Amplifiers," *AIEE Special Publication T-98*, Aug. 1957, pp. 187–204.

13. Self-Balancing Magnetic Servo Amplifiers, W. A. Geyger. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 11, Oct. 1955, pp. 987–1000.

14. Magnetic-Amplifier-Operated Ink Recorders, W. A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Sept. 1958, pp. 457–71.

15. A Miniaturized D-C Instrument Transformer, W. A. Geyger. "Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication T-116*, 1959, pp. 130–53.

16. Magnetic-Amplifier Circuits (book), W. A. Geyger. McGraw-Hill Book Company, Inc., New York, N. Y., second edition 1957, pp. 27–28, Fig. 2.3(f).

17. The Ferrotracer, A Simple Dynamic B-H Loop Tracer with Silicon-Diode Chopper Circuits, W. A. Geyger. *Proceedings*, National Electronics Conference, vol. 12, Oct. 1956, pp. 83–97.

18. The Effective Feedback Factor of Self-Balancing Magnetic Amplifiers, W. A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Mar 1958, pp. 66–69.

19. Magnetic Amplifiers (book), H. F. Storm. John Wiley & Sons, Inc., New York, N. Y., 1955, pp. 77–79, 215–19.

20. Magnetic Amplifiers with Orthonol Tape Cores, W. A. Geyger. *Proceedings*, National Electronics Conference, vol. 6, Oct. 1950, pp. 59–67.

21. A Switching Transistor D-C to A-C Converter Having an Output Frequency Proportional to the D-C Input Voltage, G. H. Royer. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, July 1955, pp. 322–26.

22. A New Self-Excited Square-Wave Transistor Power Oscillator, G. C. Uchrin, W. O. Taylor. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 43, Jan. 1955, p. 99.

23. A Variable-Frequency Magnetic-Coupled Multivibrator, R. L. Van Allen. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, July 1955, pp. 356–61.

24. Magnetic Inverter Uses Tubes or Transistors, C. H. R. Campling. *Electronics*, New York, N. Y., vol. 31, no. 11, Mar. 1958, pp. 158–61.

25. Frequency Control of Magnetic Multivibrators, W. A. Geyger. *Ibid.*, vol. 32, no. 30, July 1959, pp. 54–56.

26. Recent Developments on Magnetic-Coupled Multivibrators, William A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, May 1960, pp. 106–12.

A Saturable-Core Modulation Integrator

R. C. BARKER
ASSOCIATE MEMBER AIEE

A. J. GRUODIS
NONMEMBER AIEE

IN SIGNAL TRANSMISSION systems, employing suppressed-carrier modulated signals, it is often desirable to operate on the modulation in order to stabilize the system or alter its performance. The circuit discussed in this paper, an illustration of techniques employing saturable-core reactors that may be used for operating directly on the modulated carrier,[1] is based on the ability of a saturable-core reactor to convert a voltage integral to a magnetic flux and store it until called for. The experimental circuit is an integrator with a direct voltage gain of approximately 100 and a time constant of 4 seconds.

The quantity that is to be integrated is the modulation on a carrier, which we represent as a discrete function, equal in each half-cycle to the integral of the modulated wave during that half-cycle, as shown in Fig. 1. The discrete function is a mathematical model of a physical quantity, which must be used with care at modulation frequencies near the carrier. In the case of the integrator, which is intentionally designed to have a very low bandwidth, this limitation is usually of minor consequence.

The operations performed on the carrier are shown in Fig. 1. Fig. 1(a) shows a suppressed-carrier wave, which is converted to the half-wave signal shown in Fig. 1(b). This wave is then integrated, or accumulated, by the circuit. Fig. 1(c) shows the discrete function representing the accumulation of the signal in Fig. 1(b), with the initial condition assumed to be zero. Fig. 1(d) is the actual output waveform of the accumulator. It should be noted that the circuit discussed here accumulates voltage integrals in the form of flux increments, and delivers output on every other half-cycle.

Functional Description of Circuit

The conception and construction of the accumulator follows the scheme shown in the simplified functional block diagram of Fig. 2(a). The simplified circuit is

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted August 8, 1960; made available for printing September 21, 1960.

R. C. Barker is with Yale University, New Haven, Conn., and A. J. Gruodis is with International Business Machines Corporation, Poughkeepsie, N. Y.

The authors wish to acknowledge that the work reported in this paper was supported by the United States Air Force under Contract *AF 33 (616)-5018;* Lt. R. Van Wechel of the Experimental Systems Laboratory, WCLCXE, was the initiating officer.

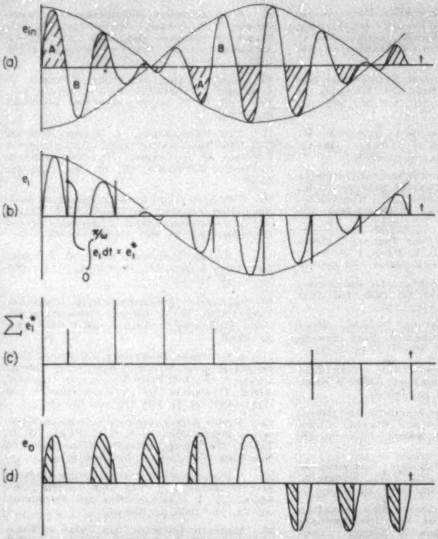


Fig. 1. (a) Suppressed-carrier input voltage; (b) input passed through gating circuit, A; (c) accumulation of integrals in (b); (d) output of accumulator

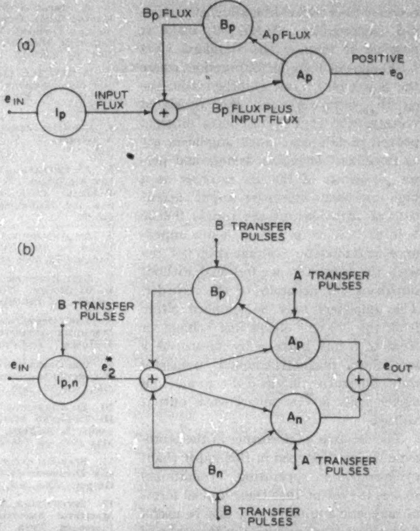


Fig. 2. (a) Functional schematic of accumulator for positive input and output; (b) functional schematic of complete accumulator

adequate to handle the type of signals shown in Fig. 1, where the integral of the input remains positive. As the figure shows, the circuit operates sequentially, there being two types of operations which take place alternately, each taking place within the fundamental time interval of one half-cycle of the carrier. The program of these two operations (identified by the *A* and *B* half-cycles shown in Fig. 1) is stored inherently in the logic of the windings and circuit connections. The two performances illustrated in Fig. 2 (a) and (b) are outlined herewith:

(a). Input is measured and stored as a flux in the positive input core I_p. Accumulated flux is transferred from positive accumulator core A_p to positive accumulator core B_p, and is simultaneously transferred to the output as an induced voltage integral.

(b). Stored input from core I_p is added to the accumulated flux in core B_p and transferred to core A_p as the new accumulated flux.

The flux in the accumulator cores is transferred around the closed transfer circuit, picking up a contribution from the input core once each cycle. This process continues while the input remains positive, the first modulation half-cycle in Fig. 1(b), or until the accumulator cores saturate. The circuit as shown, therefore, responds to positive inputs and gives positive outputs.

In order to respond to negative inputs, a second input core I_n must be added and coupled to the accumulator cores in place of core I_p to return the positive accumulated flux back to zero when the input becomes negative. Further, in order to have negative outputs, there must be a pair of negative accumulator cores—similar to the positive ones—which must also receive inputs from both input cores. The functional schematic for this circuit is shown in Fig. 2(b). It should be noted that the circuit described, although it can produce both positive and negative outputs for positive and negative inputs, is basically a half-wave circuit. The input is always measured on *A* half-cycles, and the output always appears on *A* half-cycles.

Circuit Operation

The input circuit is shown in Fig. 3. Both input cores, I_p and I_n, are pulsed

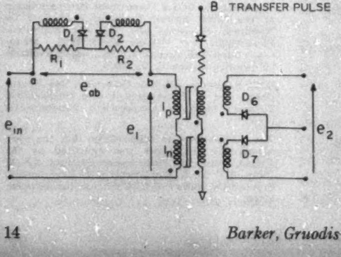
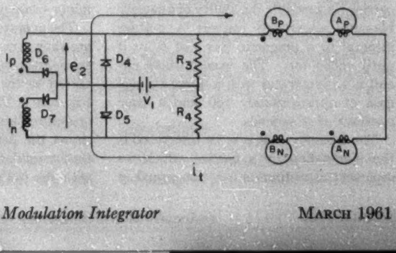


Fig. 3 (left). Partial circuit showing gating circuit, A, and input cores

Fig. 4 (right). Partial circuit showing method of inserting flux charge into flux-transfer path



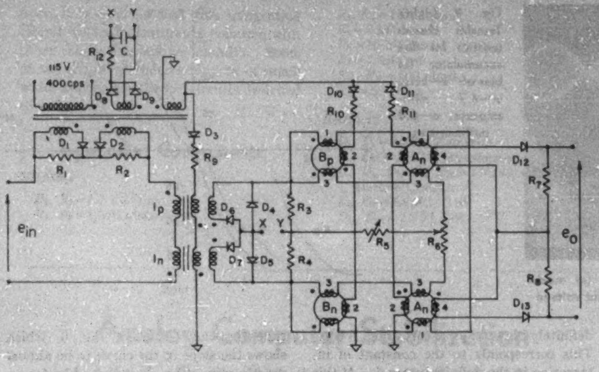


Fig. 5. Complete circuit diagram of flux-accumulator circuit

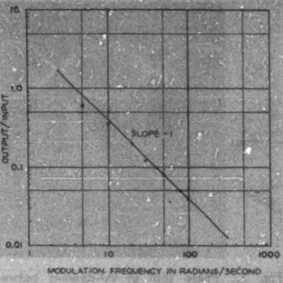


Fig. 6. Relative output of integrator circuit as a function of frequency. Line of −1 slope shown for reference

with a voltage half-sine-wave during the *B* half-cycles. The positive voltage is applied to the undotted end of the transfer winding on each input core. The cores have essentially rectangular hysteresis loops, and the negative saturation level is used as a flux reference. All of the transfer pulses used in the circuit are of sufficient magnitude to return their associated cores to negative saturation. Thus, to put a flux charge into one of the input cores requires that a positive voltage be applied to the dotted end of the input winding. The *A* gating circuit, shown connected between points *a* and *b* in Fig. 3, is powered by two windings on a reference transformer. It effectively closes the circuit between points *a* and *b* on *A* half-cycles, and opens it on *B* half-cycles, converting the voltage e_{in} of Fig. 1(a) to the voltage e_1 of Fig. 1(b). Depending upon whether the input voltage is positive or negative, its integral appears as a flux charge on either the positive or negative input core. During the *B* half-cycle, the *B* transfer pulse transfers the stored flux charge into the integral of the voltage e_2.

The remainder of the input circuit is shown in Fig. 4. The common circuit linking all four accumulator cores is called the flux-transfer circuit. The circuit to the left of and including R_3 and R_4 is designed to appear as a very low resistance on *A* half-cycles and to insert a flux increment from the input cores on the *B* half-cycles. The need for a low resistance in the flux-transfer path is the same as is found in any of the large class of circuits in which flux is to be transferred from one core to another. For example, on an *A* half-cycle, in which a flux charge is to be transferred from core A_p to core B_p, the integrated voltage equation for the flux-transfer path may be written

$$N_T\phi_A = N_T\phi_B + \int_0^{\pi/\omega} i_T R_T dT \qquad (1)$$

where i_T is the field-producing current required to move the flux charge ϕ_B into core B_p, and R_T is the total effective resistance of the flux-transfer path. The integral of the resistive term represents the flux linkages lost in the transfer process. In the circuit design, this quantity must be minimized.

The function of the voltage V_1 is to set up bias currents, determined by resistors R_3 and R_4, in the rectifiers D_4 and D_5. These bias currents are designed to be substantially greater than the maximum flux-transfer current i_T. The rectifiers D_6 and D_7 are chosen to have higher threshold voltages so that they do not carry a significant portion of the bias currents. In this manner, a low-impedance path is provided for the flux-transfer current during the *A* half-cycles in which the transfer windings of the input cores are effectively decoupled from the circuit.

On the *B* half-cycles, a *B* transfer pulse is added and applied to the input cores. If it is assumed that a flux charge were stored in the positive input core I_p during the previous *A* half-cycle, the voltage induced in the transfer winding of core I_p is of the polarity necessary to turn on D_6 and turn off D_4. The path through which the upper bias current and the flux-transfer current flow is thus switched over through D_6 so as to insert the transfer winding of I_p in series with the flux-transfer path. If there is a flux charge in core B_p at the beginning of the half-cycle, the flux charge removed from core I_p adds to it and the sum is transferred to core A_p. If, on the other hand, there is a flux charge in core B_n at the beginning of the half-cycle, the flux charge removed from core I_p subtracts from it and the difference is passed over to core A_n. When the circuit is properly adjusted, there is at the beginning of the *B* half-cycle a flux charge in one or the other of the *B* cores, but not both, and no flux charge in either of the *A* cores. If the flux charge is in the negative input core, the circuit operates in the same manner with the functions of the two halves of the input circuit reversed.

In order to trace a full cycle of events, assume that there is a flux charge stored in core A_p at the beginning of an *A* half-cycle. Also assume that there is no flux stored in core A_n. When the *A* transfer pulse is applied (a half-sinusoid from the power supply), the undotted end of the transfer winding on A_p becomes positive with respect to the dotted end. If it is assumed that the source impedance of the input circuit is zero during this interval, the voltage induced in the transfer winding of core A_p is applied directly to the transfer windings of the other three cores in series. The polarity in which the two negative cores are connected shows that the effect is to drive them further into negative saturation. The polarity of the transfer winding on B_p, however, is such that the induced voltage in A_p drives its flux in a positive direction. Thus, if it is assumed that there is no flux loss in the form of resistive voltage in the transfer circuit, the flux discharged from core A_p appears at the end of the half-cycle as a flux charge in core B_p. Thus, if the flux charge had been in core A_n instead of A_p, the flux would have been transferred to B_n instead of B_p, and the polarity of the current I_t would have been reversed.

At the end of the *A* half-cycle, then, there is a flux charge in core B_p, and a

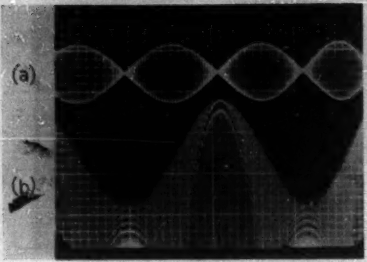


Fig. 7. Phase shift of 90-degrees between (a) modulated carrier and (b) half-cycle integrals of output voltage

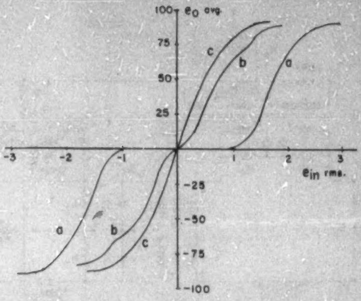


Fig. 8 (right). Transfer characteristics for flux accumulator for bias of: a—zero. b—4.7 milliamperes. c—6.0 milliamperes

flux charge representing the input voltage in either the positive or the negative input core. The other three accumulator cores are at negative saturation. When the B transfer pulse is applied, the flux charge in core B_p, augmented by the stored charge from the input cores, is transferred to A_p.

From this analysis it is apparent that, regardless of the polarity of the input voltage, the stored flux is transferred back and forth between the same pair of accumulator cores until the net accumulated flux goes to zero. Further changes in the same direction switch the accumulated flux over to the other pair of accumulator cores. The switching from one pair of cores to the other takes place automatically when the accumulated flux changes polarity, and not when the input voltage changes polarity. When the accumulated flux is positive, for example, and the input is supplied from the negative input core, the flux charge reaching core B becomes smaller on each successive cycle. Eventually the point is reached where the flux charge in core B is less than the incremental amount given by the negative input core. Thus, when the B transfer pulse is applied, the net flux charge transferred reverses polarity, the direction of the transfer current reverses, and the net flux charge is fed into core A_n instead of A_p. On succeeding half-cycles it is transferred to core B_n, and so on.

Design Consideration

In the particular circuit presented in this paper, ordinary precautions were taken to obtain a reasonable design, but no special effort was made to meet any particular specifications. However, there are some fairly obvious points to be made about the winding design. When there is no input to the circuit, it is desirable to transfer the flux charge remaining in the accumulator cores back and forth indefinitely between an A core and a B core. This corresponds to the constant of integration in the definite integral. If this can be accomplished, the result is likely to be a perfect integrator. In most physical systems using single-valued coordinates, however, there are restoring forces that result in a finite time constant and a limited range of output for the device. In this circuit, the effective restoring force is the flux-charge loss that occurs in each flux-transfer process, resulting from the integral in equation 1. The loss can be minimized, however, by reducing the winding resistances, using as much of the window space as possible with a winding of the highest possible space factor. The loss can be further reduced by supplying a bias current in separate windings, corresponding to the d-c coercive force of the accumulator cores so that no more current need flow in the transfer path than is absolutely necessary. The bias winding shown in the final circuit, Fig. 5, uses some of the rectifier bias current and has the same number of turns per core as the flux-transfer windings. This eliminates undesirable circulating currents in the bias circuit. The size of the core limits its flux linkage capacity, and thus the linear range of the circuit.

Because it is desirable to have an output from the circuit that is proportional to the accumulated flux in the cores, the output is transformer coupled. In the circuit shown, the output is supplied on every other half-cycle. To make it full-wave, output windings can be placed on the B cores as well, but some care must be taken in mixing output voltages in the load.

Experimental Performance

A diagram of the complete circuit is given in Fig. 5; the components are named in the Appendix. Performance of the circuit is best summarized by the amplitude-frequency diagram, Fig. 6, which shows the slope of the curve to be almost exactly minus one, as it should be for an integrator. To make measurements outside the frequency range shown is difficult because of the nature of the signals. Integrated output and suppressed-carrier input, shown in Fig. 7, corroborate the phase shift part of the transfer function. The time constant of the accumulator, measured near the center of its transfer characteristic, is approximately 4 seconds, resulting in a bandwidth, or "break frequency" of 0.25 radian, or 0.04 cycle per second. Direct-current transfer characteristics for three values of bias are shown in Fig. 8.

Conclusions

The method described, in which the modulation of a suppressed-carrier wave is represented by the half-cycle integrals of the wave, suggested storing and accumulating flux in saturable magnetic cores in order to integrate the modulation directly. The method is reminiscent of a differentiating scheme proposed by Scorgie,[2] which has also been developed into a form suitable for use with suppressed-carrier waves.[1] Upon analysis of errors and the mechanism of flux transfer, how-

Table I. Cores and Windings

Turns	High-Frequency Wire Size
Input Cores, Magnetics, Inc., 50061-2A	
N_1 = 1,500	30
N_2 = 1,000	33
N_3 = 220	30
A Cores, Magnetics, Inc., 50090-2A	
N_1 = 400	23
N_2 = 180	30
N_3 = 400	30
N_4 = 360	30
B Cores, Magnetics, Inc., 50090-2A	
N_1 = 400	23
N_2 = 180	30
N_3 = 400	30

ever, it is obvious that this integrating circuit has many features in common with other magnetic amplifier circuits. Some of the coupling schemes may be of much wider application than in circuits designed to integrate.

Appendix. Components

Resistors

R_1, R_2 = 4.7 kilohms
R_3, R_4 = 1,360 ohms
R_5, R_7 = 15 kilohms
R_6 = 3 kilohms
R_8 = 15 kilohms
R_9 = 1 kilohm
R_{10}, R_{11} = 500 ohms
R_{12} = 50 ohms

Rectifiers

D_1, D_2 = *1N38*, Kemtron
D_3, D_{11} = *TM12*, transistron
D_4, D_5, D_8, D_9, D_{10} = *1N153*, GE (General Electric Company)
D_6, D_7 = *1N368*, GE
D_{12}, D_{13} = *1N93*, GE

Capacitor

C = 30 microfarads

Cores and Windings

See Table I.

References

1. MAGNETIC PHASE CORRECTION CIRCUITS, R. C. Barker, A. J. Gruodis. *WADC Technical Report 59-237, ASTIA Document no. AD 214885*, Office of Technical Services, U. S. Department of Commerce, Washington, D. C., 1959.

2. LEAD NETWORK UTILIZING A SATURABLE CORE MEMORY, D. G. Scorgie. *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 71, Nov. 1952, pp. 297–302.

Analog Computer Stabilization of High-Speed Magnetic Amplifiers

W. J. GREENE
ASSOCIATE MEMBER AIEE

E. A. SAIBEL
NONMEMBER AIEE

THE 3-PHASE full-wave Ramey-type magnetic amplifier is far superior to its single-phase counterpart in band-pass characteristics.[1] In the single-phase unit, phase detection effects which produce beats in the output begin to occur with control frequencies of two-thirds of the carrier frequency. In a properly balanced 3-phase system, phase detection does not occur until the control signal frequency is raised above twice the carrier frequency.

The excellent high-frequency response of the 3-phase Ramey-type amplifier is seldom realized when it is employed in closed-loop control systems due to the inherent delay or "dead time" associated with magnetic amplifiers. The problem of retaining stability without sacrificing frequency response or precision in the control function can be solved by means of an analog computing technique called dead-time simulation. With this stabilizing procedure, the feedback signal is advanced through a phase angle corresponding to dead time, thus producing stable Nyquist plots.

The method is particularly effective when applied to Ramey-type magnetic amplifiers supplying current under rapid function control to nonlinear load impedances. This feature was exploited in developing a power source for spot welding thin sheet by electric arc methods. In this application the current must be raised suddenly to a preheat level, then gradually raised to a welding level, followed by reduction to a post-heat level, and finally tapered off to zero for crater fill. The entire program may be completed in a period of a few tenths of a second. Accurate reproduction of the function requires a band pass from zero to 80 cps (cycles per second). The electric arc is not only a nonlinear load impedance but it also exhibits a negative resistance characteristic in the low-current range. Conventional stabilization procedures used with arc loads require the insertion of added series resistance in the load circuit to produce a net positive load circuit resistance. Dead-time simulation results in stable operation even when the net load circuit resistance is negative.

Principles Employed in Dead-Time Simulation

The approximate block diagram representation of a 60-cycle Ramey-type magnetic amplifier with a load consisting of a series linear inductance and resistance is shown in Fig. 1(A). The open-loop Nyquist plot of this system at the feedback point across the resistor takes the form of a spiral which encircles the origin once for each 120 cycles of the control frequency. The 180-degree phase shift point occurs at a frequency between 30 and 60 cycles, depending on the load circuit time constant. Stabilization in closed loop is usually achieved by inserting a tuned *R-C* (resistance-capacitance) rejection filter in the feedback loop to reduce open-loop gain below unity in the region of the frequency corresponding to a 180-degree phase shift. This procedure is not useful for varying or nonlinear load impedances because the filter must be retuned for each condition of the load.

A simple analog type of feedback is shown in block diagram form in Fig. 1(B). This circuit employs a separate amplifier with the same gain as the magnetic

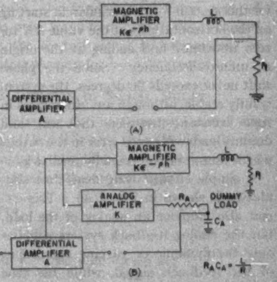


Fig. 1. Conventional and analog feedback

A–Conventional: open-loop gain to feedback point

$$G_o=AK\frac{R}{R+pL}\epsilon^{-ph}$$

Closed-loop gain

$$G_c=AK\frac{R}{(1+AK\epsilon^{-ph})R+pL}\epsilon^{-ph}$$

B—Analog: open-loop gain to feedback point

$$G_o=AK\frac{R}{R+pL}$$

Closed-loop gain across load R

$$G_c=AK\frac{R}{(1+AK)R+pL}\epsilon^{-ph}$$

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 13, 1960; made available for printing September 21, 1960.

W. J. GREENE is with the Air Reduction Company; Inc., Murray Hill, N. J., and E. A. SAIBEL is with the Rensselaer Polytechnic Institute, Troy, N.Y.

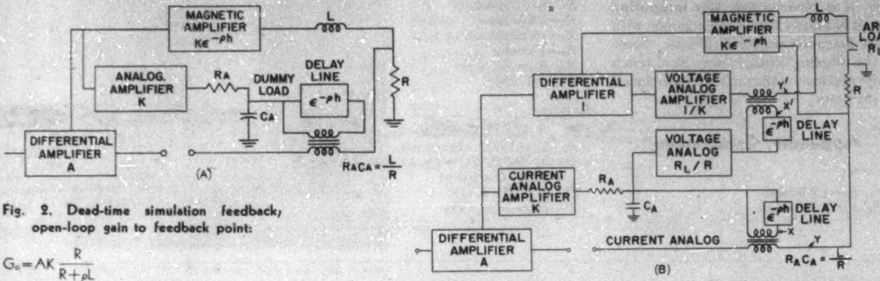

MAGNETIC AMPLIFIER Kϵ^{-ph}
L
ANALOG AMPLIFIER K
RA
DUMMY LOAD
DELAY LINE
ϵ^{-ph}
R
CA
DIFFERENTIAL AMPLIFIER A
(A)
RACA = L/R
MAGNETIC AMPLIFIER Kϵ^{-ph}
L
ARC LOAD RL
DIFFERENTIAL AMPLIFIER I
VOLTAGE ANALOG AMPLIFIER 1/K
Y'
X'
R
ϵ^{-ph}
DELAY LINE
VOLTAGE ANALOG RL/R
CURRENT ANALOG AMPLIFIER K
RA
CA
ϵ^{-ph}
DELAY LINE
X
Y
DIFFERENTIAL AMPLIFIER A
CURRENT ANALOG
(B)
RACA = L/R

Fig. 2. Dead-time simulation feedback; open-loop gain to feedback point:

$$G_o = AK\frac{R}{R+pL}$$

A—Closed-loop gain across load R

$$G_c = AK\frac{R}{(1+AK)R+pL}\epsilon^{-ph}$$

B—With nonlinear arc load: closed-loop gain across current feedback resistor R

$$G_c = AK\frac{R}{(1+AK)R+pL}\epsilon^{-ph}$$

amplifier but without the delay or dead time. A dummy load with the same time constant as the load on the magnetic amplifier provides a feedback voltage identical in magnitude with that of the real load but advanced in phase by the dead time. The open-loop Nyquist plot for this system forms a semicircle starting on the horizontal axis at the value AK for zero frequency and ending at the origin at infinite frequency. Since the phase shift never exceeds 90 degrees, the system is fully stable, although any differences in time constant between the real and dummy loads will cause error in the output which cannot be corrected by feedback.

A simple circuit using dead-time simulation is shown in Fig. 2(A). This circuit also employs an analog of the load, but the analog feedback system supplies only the dynamic or varying component of the feedback signal, while the true load circuit determines the steady-state d-c level. Consequently, a slight mismatch between the true and the dummy load will not result in a drift in output. With step function input, the analog system supplies the feedback signal only during dead time, while the true feedback signal arrives at the instant the analog signal disappears. The Nyquist plot is identical to that for the pure analog feedback system described above.

When nonlinear load impedances such as a welding arc are used with current-controlled magnetic-amplifier power supplies, a more complex circuit with both voltage and current feedback is needed for dead-time simulation. The block diagram for this type of control is shown in Fig. 2(B). The dummy load must have the same time constant as the inductance and current feedback resistor, L/R. The nonlinear arc load R_L is balanced out by positive voltage feedback so that the circuit reduces mathematically to the previous circuits with only linear inductance L and resistance R in the output.

The voltage feedback signal across the arc load is delayed by dead time in the magnetic amplifier. Consequently, an equivalent voltage produced in analog circuits without the characteristics magnetic-amplifier delay must be inserted into the voltage feedback line. This is accomplished by means of a nonlinear analog circuit which generates a voltage equal to the arc voltage corresponding to an input signal which is proportional to load current. The input signal is obtained from the dummy load rather than the current feedback resistor R to avoid introducing dead-time delay.

The voltage at the end of the delay line in both the voltage and current analog circuits (points x and x' in Fig. 2(B)) are exactly equal in magnitude and phase with the corresponding true load feedback points (y and y'). Therefore the transformer primaries may be returned to y and y' without changing the Nyquist plot or the equivalence to the simple circuit of Fig. 2(A). This connection has the advantage of using the magnetic amplifier and true load as analog components to obtain the delayed analog signals.

Improved Transform Representation

According to the simplified theory of dead-time simulation described above, it should be possible to obtain a substantially flat frequency response in closed loop to 10 or 20 times carrier frequency. Unfortunately, the gain of the magnetic amplifier will reduce to a value near zero at much lower frequencies. The error

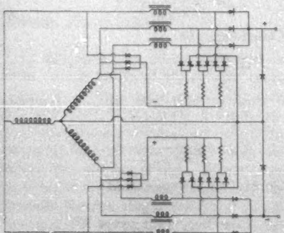


Fig. 3 (left). Load circuit of 3-phase Ramey-type magnetic amplifier

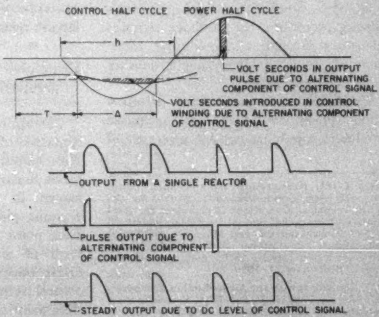

CONTROL HALF CYCLE
POWER HALF CYCLE
h
VOLT SECONDS IN OUTPUT PULSE DUE TO ALTERNATING COMPONENT OF CONTROL SIGNAL
VOLT SECONDS INTRODUCED IN CONTROL WINDING DUE TO ALTERNATING COMPONENT OF CONTROL SIGNAL
T
Δ
OUTPUT FROM A SINGLE REACTOR
PULSE OUTPUT DUE TO ALTERNATING COMPONENT OF CONTROL SIGNAL
STEADY OUTPUT DUE TO DC LEVEL OF CONTROL SIGNAL

Fig. 4 (right). Phasing of a carrier wave of angular velocity ω_0; single reactor waveforms

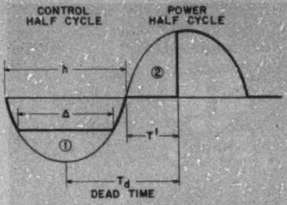


Fig. 5. Voltage in control circuit of a single reactor

in the simplified theory is caused by the deviation in the performance of the magnetic amplifier from the performance specified by the transfer function ke^{-ph} frequently employed. The assumption of constant gain and fixed delay independent of frequency is too crude for the analysis of wide-band closed-loop control systems.

Derivation of a more exact small signal gain transfer function for the full-wave 3-phase Ramey-type magnetic amplifier (see Fig. 3) is shown below.

CALCULATION OF SMALL SIGNAL GAIN

Assume that a carrier wave of angular velocity ω_0 introduced into the control and power windings of a single half-wave reactor in the Ramey 3-phase amplifier has phasing shown in Fig. 4. Also assume that a control signal of infinitesimal amplitude and angular velocity ω is superimposed on a steady d-c level with phasing as shown. The change in $\int edt$ stored in the control winding as a result of the superimposed a-c control signal is as follows:

$$\text{change in}\int edt=\int_T^{T+\Delta}\sin\omega tdt$$

Integrating and simplifying yields

$$\text{change in}\int edt=\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\left[\Delta\sin\omega\left(T+\frac{\Delta}{2}\right)\right]$$

On the conduction side of the cycle, the leading edge of the conduction phase will be delayed so that reduction in $\int edt$ of the output pulse is equal to the above change in $\int edt$ multiplied by the turns ratio N_L/N_C.

The output pulse may be considered to be the sum of a steady gating pulse caused by the control level of the direct current added to a sharp pulse of short duration equal in $\int edt$ to

$$\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\left[\Delta\sin\omega\left(T+\frac{\Delta}{2}\right)\right]$$

caused by the a-c control signal; see Fig. 4. In determining the output caused by the a-c control signal, attention can be directed only to the sharp pulse component. In a 3-phase full-wave circuit, a pulse will appear in the output at intervals of $h/3$ seconds.

Numbering the phases in sequence 1 through 6, the $\int edt$ in the output of each phase is as follows:

Phase 1:

$$\int edt=\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\Delta\left[\sin\omega\left(T+\frac{\Delta}{2}\right)\right]$$

Phase 2:

$$\int edt=\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\Delta\left[\sin\omega\left(T+\frac{\Delta}{2}+\frac{3}{h}\right)\right]$$

Phase 3:

$$\int edt=\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\Delta\left[\sin\omega\left(T+\frac{\Delta}{2}+\frac{2h}{3}\right)\right]$$

etc.

The number of pulses n for one cycle of the control frequency is equal to $6\omega_0/\omega$. If n is a positive integer, the output pulse train can be integrated over a single cycle to give the fundamental Fourier component b_1.

$$b_1=\frac{\omega}{\pi}\int_0^{\frac{2\pi}{\omega}}f(t)\sin\omega tdt$$

or

$$b_1=\frac{\omega}{\pi}\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\times$$

$$\Delta\left[\begin{array}{c}\sin^2\omega\left(T+\frac{\Delta}{2}\right)+\sin^2\omega\left(T+\frac{\Delta}{2}+\frac{h}{3}\right)+\\ \left[\sin^2\omega\left(T+\frac{\Delta}{2}+\frac{2h}{3}\right)+\right.\\ \ldots+\sin^2\omega\left(T+\frac{\Delta}{2}+\frac{(n-1)h}{3}\right)\end{array}\right]$$

if

$$n=2,\ b_1=\frac{\omega}{\pi}\Delta\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\left[2\sin^2\omega\left(T+\frac{\Delta}{2}\right)\right]$$

indicating the presence of beats since the output is a function of initial phasing of the control signal. If n is equal to or greater than 3, then

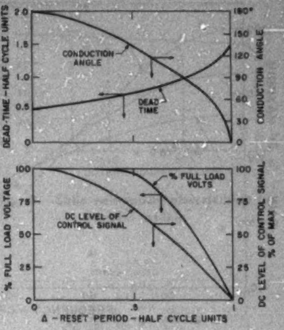


Fig. 6. Control characteristics

$$b_1=\frac{\omega}{\pi}\Delta\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\left(\frac{n}{2}\right)$$

Simplifying yields

$$b_1=\text{gain}=3\frac{\Delta}{h}\frac{N_L}{N_C}\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}$$

Hence, the fundamental component of the pulse wave form in the output caused by the control signal is independent of initial phasing for $n=3$ or greater.

CALCULATION OF SIGNAL DELAY

Referring to Fig. 5, areas 1 and 2 are equal, and it is seen that the delay in the magnetic amplifier is a function of the d-c level of the control signal.

Area 1

$$=2\int_{\frac{h-\Delta}{2}}^{\frac{\pi}{2\omega_0}}\sin\omega_0 tdt-\Delta\sin\omega_0\left(\frac{h-\Delta}{2}\right)$$

$$=\frac{2}{\pi}h\cos\omega_0\left(\frac{h-\Delta}{2}\right)-\Delta\sin\omega_0\left(\frac{h-\Delta}{2}\right)$$

Area 2

$$=\int_0^{T'}\sin\omega_0 tdt$$

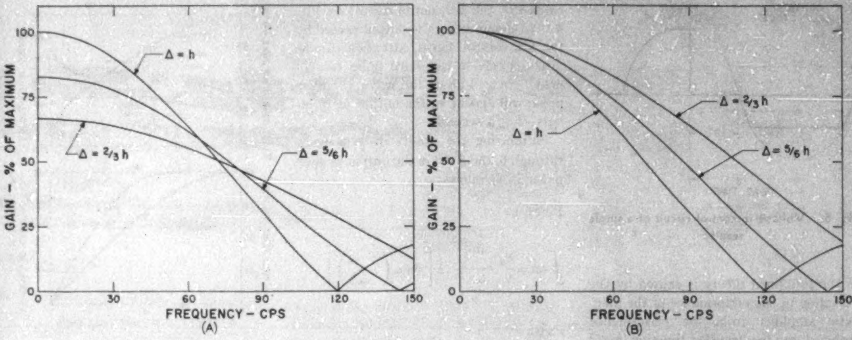
$$=\frac{h}{\pi}(1-\cos\omega_0T')$$

Equating area 1 and area 2, and simplifying, yields

$$T'=\frac{h}{\pi}\cos^{-1}\left(1-2\sin\frac{\pi\Delta}{2h}+\frac{\pi\Delta}{h}\cos\frac{\pi\Delta}{2h}\right)$$

Dead time $T_d=T'+h/2$

$$T_d=\frac{h}{2}+\frac{h}{\pi}\cos^{-1}\left(1-2\sin\frac{\pi\Delta}{2h}+\frac{\pi\Delta}{h}\cos\frac{\pi\Delta}{2h}\right)$$



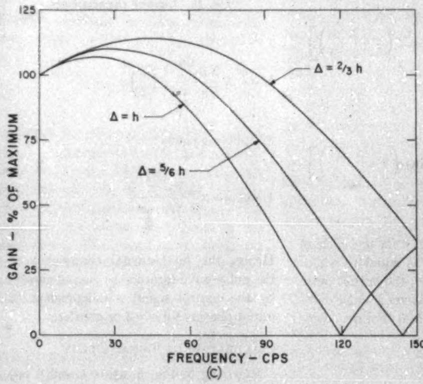


Fig. 7. Dynamic response of Ramey-type magnetic amplifier

A—Three-phase 60-cycle full-wave
B—Amplitude compensated
C—Amplitude- and frequency-compensated

Note that the magnetic-amplifier delay varies from a value of $h/2$ for large conduction angles ($\Delta \rightarrow 0$) to a value of $3/2h$ for small conduction angles ($\Delta \rightarrow h$); see Fig. 6.

In usual theoretical analysis of Ramey-type magnetic amplifiers, the delay is assumed constant at a value h. If the gain and signal delay are combined, the transfer function of the 3-phase full-wave amplifier can be represented as

$$3\,\frac{\Delta}{h}\,\frac{N_L}{N_C}\,\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}\,\epsilon^{-pT_d}$$

where

$$T_d=\left[\frac{h}{2}+\frac{h}{\pi}\cos^{-1}\left(1-2\sin\frac{\pi\Delta}{2h}+\frac{\pi\Delta}{h}\times\cos\frac{\pi\Delta}{2h}\right)\right]$$

It should be emphasized that this equation is accurate for frequencies up to twice the carrier frequency to avoid phase detection beats in the output. It is also limited to conduction angles less than 120 degrees. At greater angles of conduction, the leading edge of the output pulse will be obscured by the trailing portion of an earlier phase. Since alternating voltages in the control circuit appear in the output as a position modulation of the leading edge of the load circuit pulse, the gain would be lower than the calculated value when overlapping occurs; see Fig. 6.

The theoretical gain as a percentage of maximum gain is plotted in Fig. 7(A). Note that the d-c gain reduces with large conduction angles (Δ small) and reaches a maximum at low conduction angles ($\Delta \rightarrow h$). The frequency response curves show a series of nodes. The first node occurs at a frequency above 120 cps. The larger conduction angles show improved response at higher frequencies.

Compensation Networks for Wide-Band Magnetic Amplifiers

The nonlinear gain characteristic in the Ramey-type amplifier introduces a variation which is not present in the analog amplifiers. Although feedback in dead-time simulation control circuits will correct variation in d-c gain, the proportionality between currents and voltages in the analog and power circuits would be lost in the process. For this reason, it is preferable to employ a compensating network to linearize the over-all magnetic-amplifier gain. This is called amplitude compensation.

Amplitude compensation can be obtained by introducing a nonlinear network of resistors and diodes into the electronic amplifier driving the control circuit. The network is designed for a gain which approximates or is proportional to h/Δ over the range from $\Delta=h$ to $\Delta=2/3h$.

After amplitude compensation, the magnetic-amplifier ain gwill take the form

$$k\,\frac{\sin\frac{\omega\Delta}{2}}{\frac{\omega\Delta}{2}}$$

where k is a constant which includes the turns ratio and proportionality factor in the compensating network. Amplitude-compensated per-cent-gain curves are shown in Fig. 7(B).

In dead-time simulation the high-frequency components of the feedback signal are derived from the analog circuit rather than the load circuit. Hence, poor high-frequency response in the magnetic amplifier will not be corrected by the feedback loop. Extension of the flat portion of the gain-frequency characteristic to a point nearer to the first node requires the

Fig. 8. Function-controlled power supply

A—For arc spot welding
B—Interior view

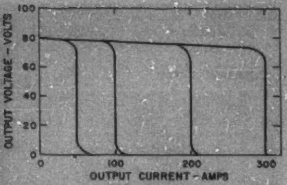


Fig. 9. Static volt-ampere characteristics for resistive loading

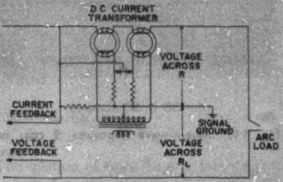


Fig. 10. Load circuit of function-controlled power supply

use of frequency compensation. More accurate reproduction of transient input signals will be obtained thereby.

A simple method for introducing frequency compensation involves the use of a dummy load which has a time constant slightly in excess of the true load. A mismatch of this kind increases the driving voltage to the control circuit of the magnetic amplifier linearly with frequency by an amount which is in excess of the requirements for load circuit inductance. The gain-frequency characteristics for an amplitude and frequency compensated amplifier are shown in Fig. 7(C).

Arc-Welding Power Supply

In the power supply designed for arc spot welding, a 3-phase 60-cycle full-wave Ramey-type amplifier is current-controlled by an electronic amplifier using dead-time simulation. The unit has a rating of 300 amperes continuous with 80 volts open circuit; see Figs. 8(A) and (B).

Open-loop gain expressed in terms of load circuit amperes per volt input is approximately 3,000. In closed loop, the gain is reduced to 20 amperes per volt. Volt-ampere curves obtained with a resistive load are shown in Fig. 9.

In the current feedback loop the magnetic amplifier itself is used to obtain the delayed analog signal. This not only prevents the 360-cycle ripple from entering the input of the high-gain electronic amplifier but also compensates for the variation in dead time with control signal level. However, in the voltage feedback loop the delayed analog signal is obtained through a delay line set for 1/2 cycle of the carrier frequency. This connection was dictated by the requirements of the welding arc load. The power supply can also be used for manual welding as well as programmed spot welding. In the manual welding application, the arc is started by direct short-circuiting of the power supply. Faster response to load circuit disturbances with the conventional delayed analog signal dictated this connection.

The circuit employed in the power supply differs in another respect from the simple circuit of Fig. 2(B). The resistance *R* required to obtain a satisfactory load circuit time constant with the minimum inductance for filtering 360-cycle ripple would introduce excessive power loss.

The load circuit was changed to obtain the effect of a series resistor by inserting a d-c current transformer connected as shown in Fig. 10. With this connection, *R* is equal in effect to the secondary volts per ampere in the primary of the d-c current transformer. Also, the load resistance R_L is the algebraic sum of arc resistance and a negative resistance equal in magnitude to *R*. The voltage analog circuit is also adjusted for this condition.

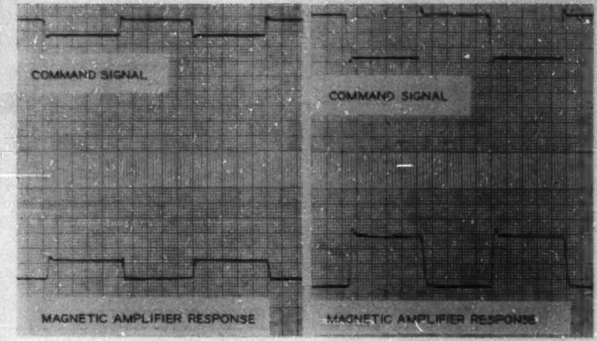
Blocking rectifiers from the output terminals of the amplifier to the center tap of the Y-connected transformer secondary were employed to prevent the flow of load current in the reactors during the reset period (see Fig. 3). This feature is required because of the inductive loading.

Individual set resistors on each reactor returned to a voltage twice as great as the voltage absorbed by the reactor permit the use of less expensive core materials. Blocking rectifiers are also employed in the set resistor circuits.

In the analysis of dead-time simulation it was assumed that the magnetic amplifier could develop any transient peak voltages commanded by the analog signal, particularly those required to initiate or stop the flow of current in the inductance during steeply rising or falling input command signals. In a practical power source, these voltages would be limited to variations from zero to open-circuit voltage. To avoid errors in correspondence between the analog and load circuits, limiters are inserted in the analog amplifiers. Analog command signals are thus prevented from overdriving the magnetic amplifier, and will permit the maximum rates of rise and fall in current consistent with the voltage available.

Cathode-ray oscillograph examination of the band-pass characteristics indicated satisfactory agreement with theory for signals of low frequency and for those of high frequency and low amplitude. The action of the limiters reduces response below the theoretical point for large amplitude signals at the upper end of the pass band.

Response to square waves of approximately 2 cps are shown in Fig. 11. Note that the distortion introduced by the pen-type recorder (80-cycle band pass)



A—30 to 90 amperes

B—1 to 180 amperes

Fig. 11. Square-wave responses, 2 cps; recorded by pen-type instrument with 60-cycle bandwidth

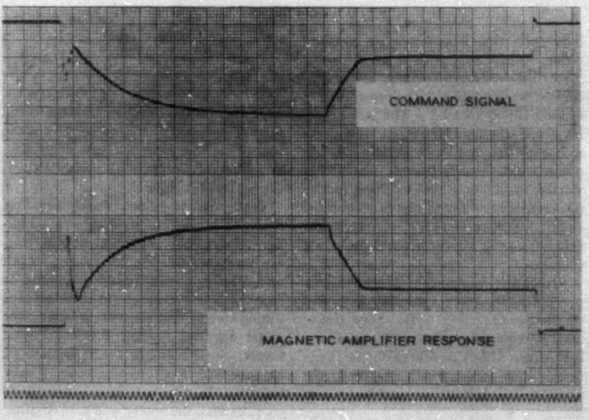


Fig. 12. Typical arc spot welding program: zero to 300 amperes

obscures any exact analysis of the rate of rise when it operates over a range from 30 to 90 amperes. Over a greater range (1 to 180 amperes) the effect of the limiters in reducing the rate of fall in the current is noticeable.

Fig. 12 shows a typical programming cycle for arc spot welding. The sharp spike of 2 cycles' duration at the beginning of the program is successfully reproduced.

Conclusions

The excellent high-frequency response of 3-phase Ramey-type magnetic amplifiers may be retained in closed-loop control systems containing a high degree of degenerative feedback. The use of analog elements for simulation of feedback signals during dead time results in stable operation with linear or nonlinear load impedances.

Analysis of dead-time simulation systems requires a more precise transfer function representation for the Ramey-type magnetic amplifier. A function suitable for the 3-phase full-wave type was developed from a theoretical analysis of small signal gain characteristics.

Analog techniques applied to the development of a program-controlled power supply for arc spot welding applications, result in performance characteristics which show good agreement with corresponding characteristics determined by theoretical means.

Appendix. Gain Calculations

Gain calculations for a Ramey-type amplifier with dead-time simulation feedback-controlling current in a nonlinear arc load follow:

$$G_0=(A+E_v)k\epsilon^{-\rho h}\frac{R}{R+R_L+\rho L}+Ak\frac{1}{1+\rho R_A C_A}(1-\epsilon^{-\rho h})$$

$$E_v=(A+E_v)\epsilon^{-\rho h}\frac{R_L}{R+R_L+\rho L}+A\frac{R_L}{R}\frac{1}{1+\rho R_A C_A}(1-\epsilon^{-\rho h})$$

$$E_v=A\left[\frac{\epsilon^{-\rho h}R_L+(R+R_L+\rho L)\dfrac{R_L}{R}\dfrac{1}{1+\rho R_A C_A}\times(1-\epsilon^{-\rho h})}{R+\rho L+R_L(1-\epsilon^{-\rho h})}\right]$$

$$G_0=Ak\epsilon^{-\rho h}\left\{\frac{R+R_L\dfrac{1}{1+\rho R_A C_A}(1-\epsilon^{-\rho h})}{R+\rho L+R_L(1-\epsilon^{-\rho h})}\right\}+\left(Ak\frac{1}{1+\rho R_A C_A}\right)(1-\epsilon^{-\rho h})$$

Matching the time constants of analog and load circuit yields

$$R_A C_A=\frac{L}{R}$$

$$G_0=Ak\frac{R}{R+\rho L}$$

and

$$G_C=\frac{AkR}{(1+Ak)R+\rho L}\epsilon^{-\rho h}$$

References

1. DYNAMIC BEHAVIOR OF A 3-PHASE MAGNETIC AMPLIFIER, F. J. Ellert. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, Nov. 1956, pp. 554–60.

2. 160,000-AMPERE HIGH-SPEED MAGNETIC-AMPLIFIER DESIGN, Allen B. Rosenstein. *Ibid.*, vol. 74, Mar. 1955, pp. 90–99.

3. A FAST-RESPONSE FULL-WAVE MAGNETIC AMPLIFIER, G. E. Lynn. *Ibid.*, vol. 77, Mar. 1958, pp. 37–41.

4. BLOCK-DIAGRAM METHOD OF ANALYSIS APPLIED TO THE SATURABLE REACTOR, Robert M. Hubbard. *Ibid.*, pp. 57–65.

5. MAGNETIC AMPLIFIER ANALYSIS AND APPLICATION USING BLOCK-DIAGRAM TECHNIQUES, Robert M. Hubbard. *Ibid.*, vol. 76, Nov. 1957, pp. 578–88.

6. OPERATIONAL MAGNETIC AMPLIFIERS WITH MULTIPLE CONTROL WINDINGS, T. T. Kadota, H. C. Bourne. *Ibid.*, pp. 515–20.

7. MAGNETIC AMPLIFIERS OF THE SELF-BALANCING POTENTIOMETER TYPE, W. A. Geyger. *Ibid.*, vol. 71, 1952 (Jan. 1953 section), pp. 383–95.

An Improved Model for Flux Reversal in Ni-Fe Cores

I. LELIAKOV
NONMEMBER AIEE

F. J. FRIEDLAENDER
ASSOCIATE MEMBER AIEE

THE PURPOSE of this work is to develop a quantitative model for the mechanism of flux reversal in nickel-iron (Ni-Fe) cores. The model allows the prediction of the instantaneous values of magnetic flux during reversal as a function of the applied magnetic-field intensity. Emphasis is placed upon the applicability of the results to practical engineering problems as well as the adherence of the model to the physical description of the process. This paper deals primarily with flux changes in grain-oriented 50% Ni-Fe tape-wound cores; however, the general approach is equally applicable to other alloys and types of core construction.

Qualitative Description of Model

A formulation of the problem is best initiated by a summary of some important results obtained in past investigations of flux-reversal mechanism.

The applied field intensities in the range $H_c<H<100\,H_c$ flux reversal in tape cores occurs primarily through domain wall motion, i.e., the growth of magnetic domains aligned with the field at the expense of unaligned domains. Thus, in order to describe the dynamic core behavior during flux reversal it is sufficient to determine the rate of motion and general geometry of domain walls.

The initial domain geometry in a previously saturated core consists of an array of randomly spaced, "semicylindrical" surface domains of reverse magnetization.[1]

The damping effects associated with the motion of domain walls in 50% Ni-Fe tapes are primarily eddy currents and spin relaxation damping. Eddy currents flowing within the material during flux change results in a reduction of the effective field intensity H_{ew} "seen" by the moving domain walls. Thus,

$$H_{ew}\overline{<}H_e=H_{\text{applied}}-H_c \tag{1}$$

where H_c is the coercive force of the core material. This reduction in effective field intensity is a function of the total rate of change of flux and hence is dependent upon wall position and over-all geometry.

Spin relaxation damping is independent of wall geometry and is described by the expression

$$\beta_r v_n=H_{ew} \tag{2}$$

where β_r is the damping coefficient associated with the particular core material and v_n is the wall velocity normal to the domain wall.

The flux changes considered are a large fraction of the maximum flux swing $2\phi_s$. Hence, the relatively small reversible flux changes in 50% Ni-Fe due to such effects as domain wall-surface tension will be neglected. Only irreversible flux changes are considered, i.e., no wall motion is assumed to exist after removal of the applied field.

Growth of Isolated Domain

Upon the application of a step field $\mathbf{H}=H\mathbf{a}_z$ to the isolated starting domain shown in Fig. 1, the following initial and boundary conditions are established.

For

$$\left.\begin{matrix}\theta=0\\ \theta=\pi\end{matrix}\right\}\mathbf{H}=H\mathbf{a}_z \qquad r_s\Big|_{t=0}=r_i$$

$$r<r_s,\ \mathbf{B}=+B_s\mathbf{a}_z$$

$$r>r_s,\ \mathbf{B}_s=-B_s\mathbf{a}_z$$

Thus,

$$\mathbf{B}=[B_s-2B_s u(r-r_s)]\mathbf{a}_z$$

where $u(r-r_s)$ is a positive unit step at $r=r_s$. Hence:

$$\frac{\partial\mathbf{B}}{\partial t}=-2B_s\delta(r-r_s)\frac{dr_s}{dt}\mathbf{a}_z \tag{3}$$

where $\delta(r-r_s)$ is a delta function. Domain walls are known to be of finite thickness. However, physical effects occurring within these boundary regions are taken into consideration by the use of the equivalent damping coefficient β_r. Then:

$$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t} \tag{4}$$

$$\nabla\times\mathbf{H}=\sigma\mathbf{E} \tag{5}$$

and hence

$$\nabla\times\nabla\times\mathbf{H}=-\sigma\frac{\partial\mathbf{B}}{\partial t} \tag{6}$$

and

$$-\mathbf{a}_z[\nabla^2H_z]=-\sigma\frac{\partial\mathbf{B}}{\partial t} \tag{7}$$

Since only the z-components of B and H are present, component subscripts are omitted:

$$\nabla^2H=-2\sigma B_s\delta(r-r_s)\frac{dr_s}{dt} \tag{8}$$

The solution of equation 8 is, for $r<r_s$:

$$H=H_{\text{applied}}-\sum_m\frac{4\sigma B_s}{m^2\pi}\frac{dr_s}{d_t}r_s^{(-m+1)}\times r^{m}\sin m\theta \qquad m\text{ odd} \tag{9}$$

and for $r>r_s$:

$$H=H_{\text{applied}}-\sum_m\frac{4\sigma B_s}{m^2\pi}\frac{dr_s}{dt}r_s^{(m+1)}\times r^{-m}\sin m\theta \qquad m\text{ odd} \tag{10}$$

At $r=r_s$, the effective field intensity at the domain wall:

$$H_{ew}=H_{\text{applied}}-H_c-\sum_m\frac{4\sigma B_s}{m^2\pi}\frac{dr_s}{dt}r_s\sin m\theta \tag{11}$$

Since at $t=0$ the normal velocity $v_n=dr_s/dt$, substituting the above result into equation 2 yields the equation of motion of the domain wall just after the application of the field:

$$\beta_r\frac{dr_s}{dt}+\frac{4\sigma B_s}{\pi}r_s\frac{dr_s}{dt}\frac{1}{m^2}\sin m\,\theta=H_e \qquad m\text{ odd} \tag{12}$$

The exact solution for the wall motion

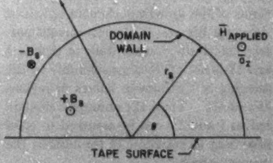


Fig. 1. Isolated surface domain

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted August 2, 1960; made available for printing September 30, 1960.

I. LELIAKOV is with the General Electric Company, Syracuse, N. Y., and F. J. FRIEDLAENDER is with Purdue University, Lafayette, Ind.

This research was supported by the National Science Foundation and the Allegheny Ludlum Steel Corporation.

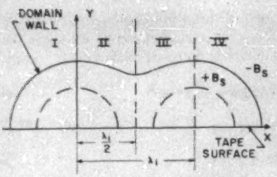


Fig. 2. Geometry of merging domains

at $t>0$ requires the knowledge of the direction normal to the wall as a function of θ and t, which results in considerable complexity of the description of domain growth. An approximation for the wall motion equation at $t>0$ can be obtained by changing equation 12 to read:[2]

$$\beta_r \frac{dr_s}{dt} + \frac{4}{\pi} \sigma B_s r_s \frac{dr_s}{dt} \theta = H_e$$

$$0 < \theta < \frac{\pi}{2} \qquad (13)$$

The error introduced near the beginning of the wall motion process is small since eddy-current damping plays only a small part near $t=0$. On the other hand, equation 13 represents a better description of wall motion for $t>0$.

Flux Reversal in Initially Saturated Tapes for High Applied Fields

The contribution to the rate of change of flux from one growing isolated surface domain is given by:

$$\frac{d\phi_u}{dt} = \int_0^{\pi/2} 4B_s r_s \frac{dr_s}{dt} d\theta \qquad (14)$$

Solving equation 13 for r_s and dr_s/dt, substituting into equation 14 and integrating gives:

$$\frac{d\phi_u}{dt} = \frac{\pi B_s H_e}{\beta_e} \log \left[\left(\frac{1}{2\beta_r}\right) \left(\frac{2\beta_e}{\beta_r} H_e t + 2\beta_e r_i + \beta_r + \sqrt{4\beta_e^2 r_i^2 + 4\beta_e H_e t + 4\beta_e \beta_r r_i + \beta_r^2} \right) \right] \qquad (15)$$

where $\beta_e = \sigma B_s$.

Flux reversal during the merging of domains is treated more conveniently in terms of the growth of domain halves as shown in Fig. 2. The eddy-current pattern across the domain walls in regions I and IV is not affected considerably by the merging of the domain halves in regions II and III. Thus, the unmerged domain halves I and IV can be assumed to grow according to equation 13 throughout the merging process.

Domain halves in regions II and III continue to grow according to equation 13 only as long as the magnetic field component produced by the eddy currents of the I and II half domains have a negligible effect on the field intensity at walls III and IV and vice versa.

At the points of maximum interaction along the domain walls II and III, the field contribution from adjoining domains reaches the values of 0.1 H_{ew} at $t=0.8\ t_{\text{merging}}$ and reaches the value H_{ew} at $t=t_{\text{merging}}$.

This interaction results in a decrease of the wall velocity in regions II and III just prior to merging. On the other hand, upon reaching the position $x=\lambda_i/2$, $y=0$, the domain walls in regions II and III undergo a rapid change in geometry due to domain wall-surface tension. A quasi-planar wall is established in regions II and III accompanied by only a small rapid change in flux $(\Delta\phi < B_s \pi r_i^2)$.

Thus, a direct transition from an unmerged to a planar geometry will represent a suitable approximation for the domain wall behavior.

The motion of the planar wall, such as in regions II and III is described by:[1]

$$H_e = 2B_s \sigma y_s \frac{dy_s}{dt} + \beta_r \frac{dy_s}{dt}$$

$$t > t_{\text{merging}} \qquad (16)$$

where y_s is the direction normal to the tape surface, and

$$y_s|_{t=0} = r_i$$

The rate of change of flux due to a pair of merged domain halves is then

$$\left[\frac{d\phi_m}{dt}\right]_i = 2B_s \frac{dy_s}{dt} \lambda_i \quad t > t_{\text{merging}} \qquad (17)$$

or

$$\left[\frac{d\phi_m}{dt}\right]_i = \frac{2B_s H_e \lambda_i}{\sqrt{\beta_r^2 + 4\beta_e(H_e t + K)}} \qquad (18)$$

where

$$K = \beta_r r_i + \beta_e r_i^2$$

At any time $0<t<t_s$ there will exist n_1 unmerged pairs of domain halves and n_2 merged pairs of domain halves, such that:

$$n_1 + n_2 = n \qquad (19)$$

where n is the number of participating surface domains per unit (pu) width of tape surface.

The spacing between domain centers is described by the probability density function $p(\lambda)$. Then, if the domain radius along the tape surface is r_m:

$$n_1 = n\,[\text{probability } (2r_m < \lambda < 1)]$$

$$= n \int_{2r_m}^{1} p(\lambda) d\lambda \quad (20)$$

and $n_2 = n - n_1$.

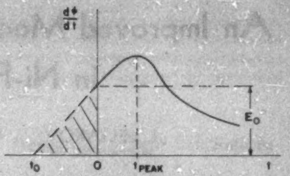


Fig. 3. Induced voltage (pu width of tape surface per turn-wrap) versus time

The total rate of change of flux pu width of tape surface is then:

$$\frac{d\phi}{dt} = n_1 \frac{d\phi_u}{dt} + \sum_i^{n_2} \left[\frac{d\phi_m}{dt}\right]_i \qquad (21)$$

For large n_2:

$$\sum_i^{n_2} \left[\frac{d\phi_m}{dt}\right]_i = n_2 \int_0^{2r_m} p(\lambda) \frac{d\phi_m}{dt} d\lambda \qquad (22)$$

and

$$\frac{d\phi}{dt} = n_1 \frac{d\phi_u}{dt} + n_2 \int_0^{2r_m} p(\lambda) \frac{d\phi_m}{dt} d\lambda \qquad (23)$$

The spacing λ between domains has a mean value $\bar{\lambda} = (1/n)$ and is restricted to the range $2r_i < \lambda < 1$. The distribution of domains appears to be best described by a Rayleigh-type probability density function:

$$p(\lambda) = \lambda 0.849^2 n^2 e^{-\frac{1}{2} n^2 0.849^2 \lambda^2} \qquad (24)$$

Since the probabilities $[0<\lambda<2r_i]$ and $[1<\lambda<\infty]$ become negligible for large n, the restriction of λ to $2r_i<\lambda<1$ may be omitted.

The reversal process described will occur for relatively large applied fields $H_e > 2H_c$. Saturation is reached for a penetration of

$$y_s|_{t=t_s} = \frac{d}{2}$$

where d is the tape thickness.

From equation 16:

$$t_s = \left[\beta_r \frac{d}{2} + \beta_e \left(\frac{d}{2}\right)^2 - \beta_r r_i - \beta_e r_i^2 \right] \frac{1}{H_e} \qquad (25)$$

Evaluation of Core Parameters

The numerical values of the core parameters r_i and β_r, as well as the number of nucleated domains, can be obtained by correlating experimental induced voltage responses with the core behavior predicted by the proposed model of domain growth.

The application of a step field $H_e > H_c$ to a tape results in the growth of n surface

Table I. Parameter Values for Tape Thickness

Tape Thickness, Mils	β_r, Ampere-Turns Second / m²	n, m 10^{-4}
0.50	17.3	2.3
1	18.5	2.2
2	19.3	2.1
4	23.6	2.5

starting domains (pu width of tape surface) with a mean radius r_i. The mean spacing between domain centers is $1/n$. Maximum induced voltage will occur approximately for maximum domain wall area, i.e., when the domains will have grown to a mean radius $1/(2n)$ along the tape surface at time t_{peak}. The motion of the walls along the tape surface is described by

$$\beta_r \left.\frac{dr_s}{dt}\right|_{\theta=0} = H_a \tag{26}$$

or

$$\beta_r r_s \Big|_{\theta=0} = H_a t_{peak} + \beta_r r_i \tag{27}$$

Hence,

$$\beta_r \frac{1}{2n} = H_a t_{peak} + \beta_r r_i \tag{28}$$

Two more relationships between core parameters can be obtained by assuming that r_i is sufficiently small to allow one to neglect the eddy-current term in the equation of motion of domain walls near $t=0$. The validity of this assumption will be considered later. Then, at $t=0$:

$$\beta_r \left.\frac{dr_s}{dt}\right|_{t=0} = H_a$$

The total rate of change of flux for n participating domains at $t = 0$ is:

$$\left.\frac{d\phi}{dt}\right|_{t=0} = 2nw\pi r_s B_s \frac{dr_s}{dt} \tag{29}$$

where w is the tape width. It will vary linearly with time near $t=0$. If the induced-voltage response to a step field H_a applied at $t=0$ is as shown in Fig. 3, the response to the field applied at $t=-t_0$ in the idealized case of $r_i=0$ would also contain the shaded area of Fig. 3. The point $t=-t_0$ is obtained by linearly extrapolating the induced-voltage curve from the value E_0 at $t=0$ to the intercept with the time axis. The values of t_0 and E_0 must satisfy the relationships.

$$r_i = \frac{H_a}{\beta_r} t_0 \tag{30}$$

and

$$E_0 t_0 = 2nw\pi r_i^2 B_s \tag{31}$$

These expressions, together with equation 28, make it possible to evaluate β_r, r_i, and n from experimentally obtained values of t_0, E_0, and t_{peak}.

The experimental measurements must be taken at relatively large fields, such that the resulting large numbers of growing domains will first merge along the tape surface. For low values of applied field, merging may first occur between domains from opposite tape surfaces which will invalidate equation 28.

In the range of experimentation in this work ($100\ H_c > H_a > 3H_c$) the calculated values of r_i and β_r have been found to remain constant. The value of r_i may be expected to increase near $H \sim H_c$ due to wall surface tension effects. Parameter values obtained for various tape thicknesses are shown in Table I.

The values of β_r obtained by this procedure compare favorably with values estimated by previous authors.[1]

The significance of the values of β_r cannot be readily discussed in terms of physical core properties, since this parameter was established as an effective macroscopic description of several physical effects. Since the flux in starting domains of reverse magnetization is relatively independent of externally applied fields, it represents a decrease from the maximum flux level theoretically possible for perfect orientation. Thus, the product nr_i may be used as a criterion for the "squareness" of the hysteresis loop of the material, i.e., perfect orientation corresponds to $r_i=0$.

The validity of the assumption made earlier, i.e., that eddy-current effects at $t=0$ may be neglected due to the small value of r_i, may now be checked by substituting mean numerical values of core parameters into the equation of wall motion 12 at $t<0$. For:

$$B_s = 1.3 \text{ weber/meter}^2$$

$$\sigma = 2.2 \times 10^6 \text{ ohm}^{-1} m^{-1}$$

$$19.6 \frac{dr_s}{dt} + 3.3 r_s \frac{dr_s}{dt} \sum_m \frac{1}{m^3} \sin m\theta = H_a$$

where maximum $r_s = r_i$.

The mean error in domain radius over $t_0 \leqslant t \leqslant 0$ is found to be 7.4%. More accurate results may be obtained by considering the quantities $d\phi/dt$ and $d^2\phi/dt^2$ at $t=0$ and including eddy-current effects. This method, however, has been found to require the use of a computer.

PREDICTION OF CORE RESPONSE

With the knowledge of the core parameters β_r, r_i, and n the instantaneous domain wall geometry and flux variation may be predicted from equations 13, 16, and 23.

Fig. 4 shows good agreement between the predicted rate of change of flux and the experimental response. A detailed description of the experimentation has been given in another of our papers.[3] Likewise, good agreement between predicted and experimental values has been obtained for response characteristics such as the initial induced voltage E_0, as shown in Fig. 5, as well as the time to

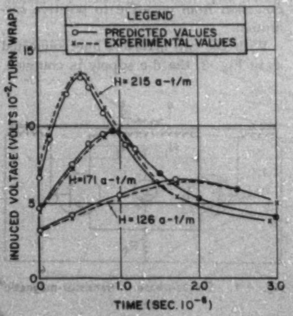


Fig. 4 (left). Typical induced voltage curves versus time

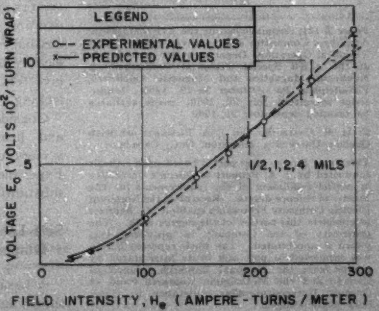


Fig. 5 (right). Initial induced voltage E_0 versus effective applied field intensity H_a

peak induced voltage and the saturation time.

Conclusions

The model developed allows the prediction of instantaneous flux variation in initially saturated cores over a wide range of field intensities and thus represents an improvement over previously proposed models of limited applicability.[1] The model is also directly applicable to the description of core behavior of initially unsaturated cores, as well as to the analysis of flux reversal for low fields ($H \sim H_c$). In the latter case, the model needs to be extended to include the domain wall geometries produced by very small numbers of participating domains.

It is recognized that the physical effects associated with the flux-reversal process have been given a simplified interpretation. However, this approach has been found necessary for the retention of reasonable simplicity in the model.

References

1. FLUX REVERSAL IN MAGNETIC AMPLIFIER CORES, **F. J. Friedlaender.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, July 1956, pp. 268–78.

2. IRREVERSIBLE FLUX CHANGES IN 50% NI-FE TAPE CORES, **I. P. Leliakov.** *Ph.D. Thesis*, Purdue University, Lafayette, Ind., June 1960.

3. INITIATION OF FLUX REVERSAL IN MAGNETIC-AMPLIFIER CORES, **F. J. Friedlaender, I. P. Leliakov.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 80, 1961 (*Paper no. 60–874*).

Self-Locking Polyphase Transistor-Magnetic Inverters

C. H. R. CAMPLING
ASSOCIATE MEMBER AIEE

J. A. BENNETT
STUDENT MEMBER AIEE

A NUMBER OF METHODS for phase-locking a group of two or more single-phase transistor-magnetic inverters have been described in the literature since 1955.[1–3] Phase displacement between inverters in most cases depends upon supplementary reactors which add to the cost and complexity of the circuit.

In the basic phase-locking arrangement described by Milnes, for example, not only are auxiliary timing reactors required but they must have associated with them auxiliary windings on the inverter cores as well as rectifiers and resistors.[1] The phase displacement depends directly upon the design of almost all of these auxiliary components, so that, while the method is quite general as well as elegant and ingenious, the difficulties which beset the designer are great.

In the 2-phase inverter described by Card, a master–slave relationship exists between the two inverters, and there is no need for supplementary reactors.[2] The slave inverter is identical with the master except that the main winding on its core is separated into two sections, and these are connected to taps on the main winding of the master inverter. The circuit is such that the slave unit is locked 90 degrees from the master. Unfortunately, this method in its simplest form is limited to quadrature phase-locking.

This paper presents a generalization of the Card method, which permits one or more slave inverters to be phase-locked to a master inverter with arbitrary phase displacements between master and slaves. A 2-phase inverter with 90-degree separation between phases is a particular case. A symmetrical 3-phase inverter with 120-degree separation between phases is another. No supplementary reactors are required. The slave inverters are driven from taps on the main winding of the master unit. Because no auxiliary circuit elements are required for locking purposes, this class of polyphase inverters is described herein as self-locking.

Consideration of the analysis, design, and loading of these self-locking inverters forms the substance of this paper. Experimental results are presented to substantiate the theory.

Self-Locking with Arbitrary Phase Displacement

A basic transistor-magnetic inverter is shown in Fig. 1. The circuit is a common-emitter arrangement. The development described in this paper need not be restricted to this form.

The frequency of the single-phase unit of Fig. 1 in cycles per second (cps) is given approximately by the equation

$$f = E/4N_1\phi_m \tag{1}$$

where E is the supply voltage, N_1 the number of turns in half of the inverter main winding, and ϕ_m the saturation flux of the core.

To illustrate the basic self-locking mechanism, consider two single-phase inverters of the type shown in Fig. 1. Let these two inverters be designed for somewhat different frequencies when operated independently from the same d-c supply voltage. This state of affairs can be realized by making the flux-turn products $N_1\phi_m$ different for the two inverters. If identical cores are used, the number of main winding turns N_1 may be different. In general, both windings and cores may differ. If the two inverters are designated A and B their frequencies, when operated from the same supply, have the ratio

$$f_A/f_B = N_{1B}\phi_{mB}/N_{1A}\phi_{mA} = \lambda_B/\lambda_A \tag{2}$$

where λ_A and λ_B are the volt-time areas characteristic of the main windings of inverters A and B, respectively, for flux excursion from negative to positive saturation.

With the two inverters interconnected as in Fig. 2, the d-c supply is common,

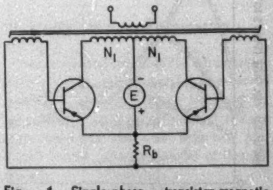


Fig. 1. Single-phase transistor-magnetic inverter

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 26, 1960; made available for printing September 21, 1960.

C. H. R. CAMPLING and J. A. BENNETT are with Queen's University, Kingston, Ont., Canada.

This paper for the most part is based upon a thesis submitted by J. A. Bennett to Queen's University in partial fulfillment of the requirements for the Master of Science degree. Receipt of the Northern Electric Company Fellowship enabled Mr. Bennett to complete this portion of the degree work. The generosity of the Company in providing this award is appreciated. The work reported herein was supported in part and in its later stages by grants from the National Research Council of Canada and the McLaughlin Research Fund of Queen's University. This assistance is gratefully acknowledged.

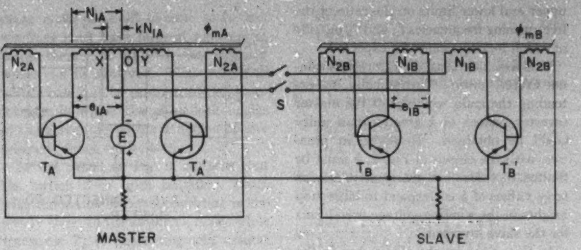


Fig. 2. Self-locking 2-phase inverter

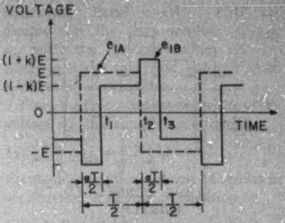


Fig. 3. Winding-voltage waveforms for 2-phase inverter

the *B*-unit is driven from taps on the *A*-unit main winding, and there can exist a master–slave relationship between the two. The tap ratio *k*, as shown in Fig. 2, is the ratio of the turns measured to the tap *X* or the tap *Y* from the center tap *O* of the *A*-unit main winding to the turns N_{1A} in half of the main winding. When the inverters are locked in this way, the common frequency is f_A, and the phase angle between them is a function of the ratio λ_B/λ_A and of the tap ratio *k*.

The steady-state waveforms of the voltages across the main windings of the master and slave inverters are shown in Fig. 3. This figure correlates with Fig. 2. The period of the master inverter is designated *T*

$$T=1/f_A \tag{3}$$

The phase displacement in seconds is $\alpha T/2$. Thus α is a numeric which is a measure of the phase displacement between master and slave. The phase displacement θ in electrical degrees is

$$\theta=180\alpha \tag{4}$$

The steady-state operation of the locked pair proceeds as follows. Between t_1 and t_2, transistor T_A of the master inverter is in the ON state as well as transistor T_B of the slave inverter. Transistors T_A' and T_B' are off. The voltage e_{1B}, impressed across the half of the slave main winding connected to T_B, is $E(1-k)$. After time t_2, when the master inverter switches, T_A is cut off and T_A' is on. Thus, in the interval from t_2 to t_3, the voltage e_{1B} is $E(1+k)$. At t_3, the slave inverter is forced to switch because its core has reached saturation. From Fig. 3, the volt-time area absorbed by the master core in one-half cycle is

$$\lambda_A=ET/2 \tag{5}$$

From the same figure, the volt-time area absorbed by the slave core in one-half cycle is

$$\lambda_B=E(1-k)(1-\alpha)T/2+\alpha E(1+k)T/2 \tag{6}$$

From equations 5 and 6, the ratio of the volt-time areas is

$$\lambda_B/\lambda_A=(1-k)(1-\alpha)+(1+k)\alpha \tag{7}$$

This ratio is also the ratio of maximum flux linkages for the two cores, i.e., the flux linkages produced by saturation flux linking all the turns ($2N_{1B}$ or $2N_{1A}$) of the main windings. In the interests of brevity, this quotient will be called the linkage ratio. The linkage ratio is defined by the equation

$$\rho=N_{1B}\phi_{mB}/N_{1A}\phi_{mA}=\lambda_B/\lambda_A \tag{8}$$

Equations 7 and 8 describe the constraints which determine behavior of the locked inverters. When equation 8 is substituted in equation 7, an equation for the fractional phase displacement α is obtained

$$\alpha=\frac{1}{2}\left(1+\frac{\rho-1}{k}\right) \tag{9}$$

From equations 9 and 4, the phase displacement in degrees is

$$\theta=90\left(1+\frac{\rho-1}{k}\right) \tag{10}$$

If the cores used in the master and slave are identical, equation 8 reduces to

$$\rho=N_{1B}/N_{1A} \tag{11}$$

In practice, such a choice of cores is desirable in most cases. However, if precise control of phase displacement is required, it should be noted that cores with identical dimensions and of the same material are likely to have noticeably different saturation fluxes. Equation 11 demonstrates that the phase displacement can be controlled effectively by changing the turns N_{1B} on the slave main windings. In fact, one way to obtain continuous control, should it be desirable, would be to make N_{1B} continuously variable by the use of some form of variable-ratio transformer technique in the construction of the slave-inverter main winding.

If the further restriction is imposed that $N_{1A}=N_{1B}$, or, more generally, if ρ is made

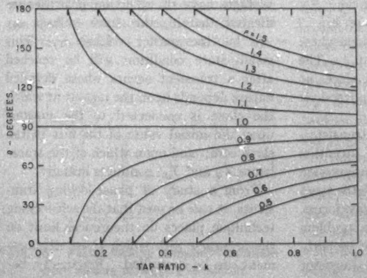
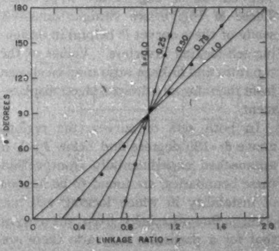


Fig. 4 (left). Phase-lock angle as a function of tap ratio (linkage ratio as parameter)

Fig. 5 (right). Phase-lock angle as a function of linkage ratio (tap ratio as parameter)



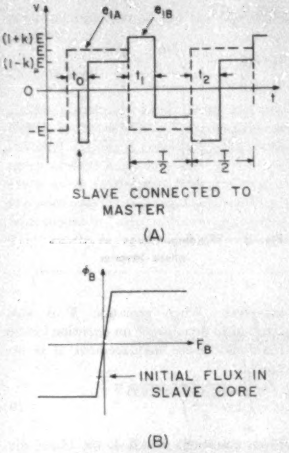


Fig. 6. Phase locking transient—slave lagging master

A—Collector-winding voltage waveforms
B—Initial condition of flux in slave core

equal to unity, the phase displacement between master and slave will be 90 degrees. The inverter used by Card to supply power to a 2-phase servomotor is an example of this special case.[2] Equation 10 also shows that, for this case, the phase displacement is independent of the tap ratio k.

Design Considerations for Self-Locking Inverters

The families of curves in Figs. 4 and 5 are derived from equation 10. They show: .

1. The effect upon the phase displacement of varying the tap ratio k when the linkage ratio ρ is taken as a parameter.
2. The effect upon the phase displacement of varying ρ when k is taken as a parameter.

The curves of Fig. 4 are hyperbolas. Those of Fig. 5 are straight lines. A study of these curves is helpful in designing self-locking inverters. Values of the tap ratio and linkage ratio may be chosen from them for any desired phase displacement.

In both sets of curves, the regions above $\theta = 180$ degrees and below $\theta = 0$, if approached experimentally from within these boundaries, are found to be regions of instability in which locking will not occur. One can infer from the curves that for a particular tap ratio there are upper and lower limits on the ratio of the free-running frequencies f_A and f_B outside of which locking cannot be expected.

It is assumed that the tap ratio k cannot exceed unity. Theoretically, by extending the main windings of the master inverter, values of k greater than unity could be obtained. However, in practice, with the circuit of Fig. 2, k must be limited to values less than unity because large values of k correspond to large step heights in the winding-voltage waveforms for the slave inverter.

In Fig. 3 the height of the step in the main-winding voltage is

$$(1+k)E-(1-k)E=2kE \qquad (12)$$

As k approaches unity, this step magnitude increases. The voltage $(1-k)\,E$ approaches 0. The feedback windings of the slave inverter have N_{2B} turns and the voltage induced in these windings has the same stepped waveform as that shown for e_{1B} an Fig. 3. Thus, the base drive voltage during the lower portion of this waveform will be $(N_{2B}/N_{1B})\ (1-k)E$. This voltage likewise will approach 0 as k approaches unity. Therefore, the drive which holds one slave transistor on and the other off is affected. If k is too large, the slave-inverter transistors will not switch correctly and proper synchronism between master and slave may be lost.

Examination of the curves of Fig. 5 shows that as k approaches 0, the phase displacement becomes critically dependent upon the linkage ratio ρ. In the limit, when $k=0$, the inverters operate independently. The slave is then free. There is no self-locking action. Reliable self-locking requires that k exceed some minimum value greater than zero.

It is difficult to specify exactly what range of values of k can be used with success, although experimental points shown in Fig. 5 give some idea of what is possible. These points were obtained from no-load tests. Diminution of slave base-drive, caused by large k, limits the slave's ability to carry load. For this reason, when the circuit of Fig. 2 is used, it is wise to choose k somewhere in the region between 0.2 and 0.3. The curves of Fig. 5 show that this places no restriction upon the range of phase-lock angles which can be obtained.

In the design of phase-locked transistor-magnetic inverters, particular attention must be directed toward the suppression of voltage spikes coincident with the steep portions of winding-voltage waveforms. Spikes of this sort are often a problem in single-phase inverters. If not properly suppressed, they may cause transistor failure. In polyphase inverters, they are likely to cause premature switching and loss of proper synchronism. In testing experimental self-locking inverters, it has been found that these spikes can be suppressed for a wide range of values of k. Suppression elements for this purpose are not shown in Fig. 2 and indeed are not always necessary.

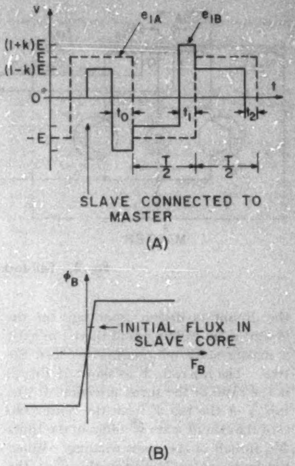


Fig. 7. Phase locking transient—slave leading master

A—Collector-winding voltage waveforms
B—Initial condition of flux in slave core

Analysis of Synchronizing Transients

No consideration is given in the foregoing analysis to the nature of the locking transients which occur when a slave inverter is connected to the master. This problem is susceptible to analysis. It can be shown for instance that, when the circuit of Fig. 2 is used and when the winding voltages have the reference polarities indicated therein, the slave voltage e_{1B} will lag the master voltage e_{1A}. This steady-state condition will be reached after a transient occurs whose detailed nature depends upon the instant at which the slave is connected to the master; upon the initial value of the flux in the slave core; and upon which of the transistors T_B and T_B' conducts initially.

From a study of phase-locking transients, it can be seen that the self-locking technique places no theoretical limit on the accuracy with which phase displacement can be controlled. Practically, because cores come in discrete sizes and windings must have integral numbers of turns, there is a limit on the accuracy which can be achieved. However, with a little care it is no problem to design an inverter in which the phase-lock angles of the slaves are within a degree of any selected design objectives.

In the circuit of Fig. 2, suppose that the switch S is open initially. Then, with the master inverter operating in the steady state, let the switch be closed while transistor T_A is conducting and assume that slave transistor T_B' conducts initially. With this set of initial conditions, the flux in the slave core will decrease toward the negative saturation level.

Suppose that the initial flux condition in the slave core is such that it saturates and switching first occurs for the slave t_0 seconds after switching occurs in the master. These conditions are depicted in Fig. 6. The choices of initial flux level in the slave core and of the instant at which the slave is connected to the master are only two of the many possibilities which could be considered.

Following the interval t_0, the flux in the slave core will change—first at a rate determined by the voltage $E(1-k)$, and then after the master switches, at a faster rate determined by the voltage $E(1+k)$. At a time t_1 seconds after the master switches, the slave core will saturate in the opposite sense and the slave inverter will switch. For this half-cycle of slave operation, the volt-time equation is

$$\lambda_B=(T/2-t_0)(1-k)E+t_1(1+k)E \tag{13}$$

From equations 5 and 8, the volt-time area λ_B also can be expressed as

$$\lambda_B=\rho ET/2 \tag{14}$$

From equations 13 and 14, t_1 can be expressed in terms of t_0, and the result is

$$t_1=\frac{T}{2}\,\frac{\rho-1+k}{1+k}+t_0\,\frac{1-k}{1+k} \tag{15}$$

Combining the latter equation with equation 9, the time interval t_1 can be expressed in terms of the steady-state fractional-phase-displacement α for which the inverter is designed. The result is

$$t_1=\frac{k\alpha T}{1+k}+t_0\,\frac{1-k}{1+k} \tag{16}$$

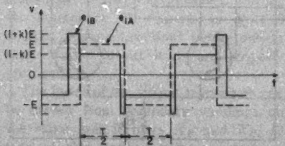


Fig. 8. Phase locking transient—transition from leading to lagging condition

A similar argument, applied to the subsequent half-cycle, leads to an expression for t_2 in terms of t_1

$$t_2=\frac{k\alpha T}{1+k}+t_1\,\frac{1-k}{1+k} \tag{17}$$

Substitution of equation 16 in equation 17 yields

$$t_2=\frac{k\alpha T}{1+k}+\frac{k\alpha T}{1+k}\,\frac{1-k}{1+k}+t_0\left[\frac{1-k}{1+k}\right]^2$$

A continuation of this argument leads to the conclusion that after n half-cycles, the displacement between master and slave in seconds is

$$t_n=\frac{k\alpha T}{1+k}\,[1+h+h^2+\ldots+h^{n-1}]+t_0h^n$$

where

$$h=\frac{1-k}{1+k}$$

As n grows large, the effect of the second term becomes negligible and the series in the first term converges to a limit since h is less than unity

$$t_\infty=\lim_{n\to\infty} t_n=\alpha T/2 \tag{18}$$

The analysis leads to the conclusion that, at least for initial conditions of the type chosen, the slave approaches a steady-state condition in which it lags the master by $\alpha T/2$ seconds. The phase-locking transient is not characterized by an overshoot.

A second set of initial conditions could be postulated in which the slave inverter is connected to the master, the latter being in the steady state, in such a way and at such a time that the slave winding-voltage e_{1B} initially leads the master winding-voltage e_{1A}.

Suppose that the slave inverter is connected to the master shortly after the voltage e_{1A} has become positive, transistor T_A conducting. Let the initial flux level in the slave core be positive as shown in Fig. 7(B). Suppose further that transistor T_B in the slave inverter conducts initially. In this case, the flux in the slave core will increase initially toward the positive saturation level. Actually T_B' is more likely to conduct initially, because the initial emitter-collector voltage is greater for it than for T_B. However, should T_B conduct first, the sequence of events depicted in Fig. 7(A) will take place. The slave will switch as soon as its core saturates. Thereafter, switching in the slave unit will proceed in the usual way, i.e., the flux in the slave core will swing between positive and negative saturation. However, Fig. 7(A) indicates that the slave winding voltage e_{1B} leads the master winding voltage. Analysis shows that this leading phase relationship is not stable.

From the first complete half-cycle of the voltage e_{1B}, this equation is obtained

$$\lambda_B=(T/2-t_1)(1-k)E+t_0(1+k)E \tag{19}$$

Combining equations 19 and 14 results in an equation analogous to equation 16

$$t_1=-\frac{k\alpha T}{1-k}+t_0\,\frac{1+k}{1-k} \tag{20}$$

A similar relationship exists between t_2 and t_1 from which, together with equation 20, it follows that

$$t_2=-\frac{k\alpha T}{1-k}(1+1/h)+t_0/h^2$$

Continuing the argument

$$t_n=-\frac{k\alpha T}{1-k}[1+h^{-1}+h^{-2}+\ldots+h^{-(n-1)}]+t_0h^{-n}$$

This result can be expressed also in the form

$$t_n=\alpha T/2+(t_0-\alpha T/2)h^{-n} \tag{21}$$

Equations 21 and 18 should be compared. Unless t_0 is exactly equal to $\alpha T/2$, an unstable condition, the phase displacement t_n in seconds will approach either 0 or $T/2$, depending upon whether t_0 is less than or greater than $\alpha T/2$, whereupon, if t_n approaches 0 for example, a transition of the form shown in Fig. 8 will take place. Thereafter, the form of analysis presented in equations 13 through 18 applies.

These transient analyses show that, for the circuit of Fig. 2, the voltage e_{1B} must lag the voltage e_{1A} in the steady state by the angle θ given in equation 10. Depending upon the design of the inverter, the angle θ can have any value between 0 and 180 degrees. These statements apply specifically to the connections and reference voltage polarities shown in Fig. 2. They do not imply that leading phase angles are unobtainable. For example, in order to get a 3-phase inverter with 120-degree separation between phases, the slaves are designed in accord-

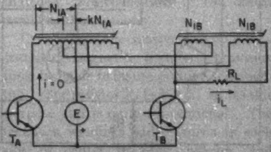


Fig. 9. Collector-to-collector loading of slave inverter

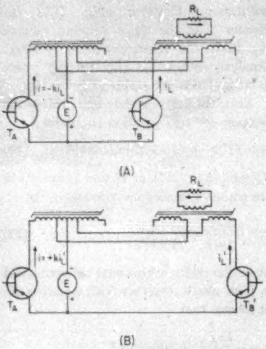


Fig. 10. Load connected to winding on slave core

A—Conditions during initial portion of master half-cycle

B—Conditions during final portion of master half-cycle

ance with the preceding theory for phase displacements of 60 degrees and 120 degrees. Reversal of the output leads, or the tap points, of the 60-degree slave will result in a balanced 3-phase inverter with one slave leading and the other lagging the master by 120 degrees.

Examination of Loading Methods

The winding voltages for the slave inverter in a self-locking system are not rectangular as is the collector-to-collector voltage. This point was noted by Card in his description of a 2-phase inverter with 90-degree displacement between phases $(\rho=1,\ \alpha=1/2)$.[2] This voltage condition holds true in general for self-locking inverters, regardless of the value of the linkage ratio ρ, provided that the taps on the master-inverter main winding through which the slave is driven are displaced from the center tap by equal numbers of turns.

If a polyphase load is to be supplied from a self-locking inverter in such a way that the phase-voltage waveforms are rectangular, the phase loads driven by the slaves can be connected directly from collector to collector. Some features characteristic of this arrangement are:

1. The phase loads must be electrically separate.

2. The output voltage will have an amplitude equal to twice the d-c supply voltage E. The range of supply voltages with which transistor-magnetic inverters can be used is limited to relatively low values by the voltage ratings of present-day available transistors. Higher output voltages can be obtained by extending the main windings of the transformers beyond the collector terminals by equal numbers of turns. With this autotransformer arrangement, however, the waveform of the load voltage will be stepped rather than rectangular.

3. When collector-to-collector loading is used with a self-locking inverter, the slave loads do not load the master. That is, load current in a slave load produces no net mmf (magnetomotive force) in the master core and none of the slave load current flows through the transistors in the master inverter.

To demonstrate the third point, consider the circuit of Fig. 2 when a resistive load is connected collector-to-collector to the slave. Suppose that transistor T_A is on in the master and T_A' off. The essentials of the circuit during an interval when these conditions hold are shown in Fig. 9. Consideration of mmf in the transformer cores indicates that the load current in each half of the slave-transformer main winding will be i_L and the load component of collector current for transistor T_B will be $2i_L$. If the taps on the master transformer which feed the slave are equidistant from the center tap, then the turns kN_{1A} on each side of the center tap must carry current i_L. The load component of supply current will be $2i_L$ and the load component of collector current i in transistor T_A must be 0. Similar analysis applies for other combinations of ON transistors, e.g., T_A on and T_B' on.

If load is connected to a separate output winding on a slave inverter, the load-voltage waveform will be stepped rather than rectangular and the ON transistor in the master inverter will carry a component of current caused by the slave's load. Nevertheless, self-locking inverters can carry load in this manner if designed to do so. Load connected to any single winding on the slave core falls within this category. Both the magnetizing current for the slave core and the base-drive current for the slave ON transistor can be classed as loads of this type.

Refer again to the circuit of Fig. 2 and consider an interval when T_A is conducting in the master and T_B in the slave. Suppose a load resistance R_L is connected across an output winding on the slave core. During such as interval, the essentials of the circuit of Fig. 2 are shown in Fig. 10(A). Transistor T_B carries the load current i_L. As far as the master is concerned, the loading effect is the same

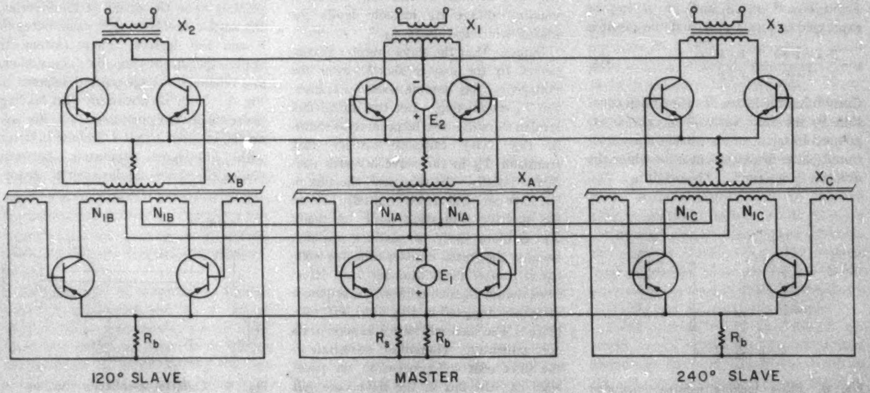


Fig. 11. Self-locking 3-phase power inverter

as if a load resistance were connected between the left-hand tap and the positive terminal of the d-c supply. In order that the net mmf in the master core be 0, the load component i of collector current for transistor T_A must be $-ki_L$. If the master itself is loaded, then the slave loading during this interval causes a reduction in the load current carried by T_A. If the master has no load, the current carried by T_A during this interval actually will be negative provided the current ki_L is greater than the no-load current carried by the collector of T_A.

The circuit of Fig. 10(B) shows the conditions which exist after the slave switches. The load current reverses. Its magnitude becomes $i_L(1+k)/(1-k)=i_L'$. This reversal, in turn, causes a reversal of the collector current i in transistor T_A, which now becomes $+ki_L'$ and augments any component of collector current in T_A caused by load on the master phase. This reasoning shows, and experiments confirm, that the self-locking inverter will carry balanced or unbalanced loads within reasonable limits, even when these loads are fed from secondary windings on each core.

In applications where any or all of the foregoing features are considered to be limitations, a self-locking inverter can be used as a timing system to control the switching of a set of power inverters. This general attack has been described by Jewett and Schmidt.[4] However, the timing inverter used by these authors was the same type as described by Milnes.[1]

As an example of this technique, the schematic for a 3-phase power inverter is shown in Fig. 11. For 120-degree separation between phases, providing that the three cores A, B, and C are identical, the number of turns required on the halves of the two slave main windings are, from equations 10 and 11

$$N_{1B}=N_{1A}(1+k/3) \qquad (22)$$

and

$$N_{1C}=N_{1A}(1-k/3) \qquad (23)$$

where N_{1A} is the number of turns on half the main winding of the master and k is the tap ratio. Slave B will lag the master by 120 degrees. Slave C will lag the master by 60 degrees—or by 240 degrees if the connections to the master taps or the output leads are properly arranged.

The bases of the output transistors can be driven from secondary windings on the slave transformers X_B and X_C, as long as the stepped voltage waveform so obtained and the loading effect of the slaves on the master transistors of the timing circuit are taken into account in the system design.

The arrangement shown in Fig. 11 has several advantages. The high-quality toroid cores which are desirable in the phase-locking circuit need not be large. Likewise, the transistors in this portion of the circuit need not have large ratings. The transformers in the output stages, however, can be built with low-cost cores of ordinary transformer steel and their construction can be considerably more economical. They should be designed to operate without saturating.

The output windings of X_1, X_2, and X_3 can be Y-connected if desired, but a delta connection, though commonly used for sinusoidal operation, cannot be used when the phase-voltage waveforms are rectangular.[4] For good frequency regulation, the source E_1 in the timing circuit should have its voltage regulated. Sources E_1 and E_2 can be common if desired.

Experimental 3-Phase Inverter

The first self-locking 3-phase inverter constructed to test the validity of the preceding theory was not designed for a specific application. Consequently, the choice of input voltage and output frequency was quite arbitrary, and the components used in most cases were those which happened to be at hand. The inverter frequency was about 95 cps. The circuit was similar to the self-locking portion of the inverter shown in Fig. 11. In the master and in each slave, a series R-C circuit was connected from collector to collector to suppress spikes in the winding voltages. Delco (General Motors Corporation) type *2N442* transistors were used. Considerable effort was made to secure, with maximum precision, a symmetrical 120-degree sequence.

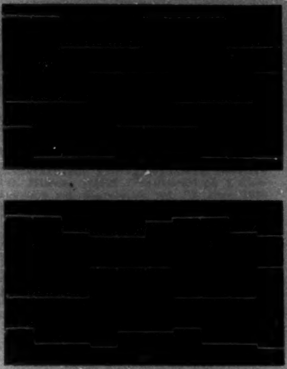


Fig. 12. Oscillograms of voltage waveforms for 3-phase inverter

A—Collector-to-collector voltages
Middle—master
Top and bottom—slaves
B—Winding voltages
Middle—master (collector-to-collector)
Top and bottom—slaves (collector winding)

If precision is required in the determination of phase angles between master and slaves, account must be taken of differences in the saturation fluxes of the cores, even though these cores have the same physical dimensions and are made of the same material. Deltamax cores (The Arnold Engineering Company no. 4180-D2) were used. The supply voltage chosen was 6 volts. These choices led to a figure of 155 turns for half of the master-inverter main winding. This winding was tapped at 39 turns on each side of the center tap, which made k nearly one-fourth. If the saturation fluxes for the cores had been identical, equations 22 and 23 indicate that the slave main windings would have required 168 turns and 142 turns, respectively. However, from the single-phase tests on the assembled inverters, it was possible to determine from equation 2 the actual ratios of saturation fluxes, which were found to be

$$\phi_{mB}/\phi_{mA}=1.05$$

$$\phi_{mC}/\phi_{mA}=1.003$$

From equations 10 and 8, the general forms of equations 22 and 23, which apply in the design of a symmetrical 3-phase inverter, are

$$N_{1B}=N_{1A}(\phi_{mA}/\phi_{mB})(1+k/3) \qquad (24)$$

$$N_{1C}=N_{1A}(\phi_{mA}/\phi_{mC})(1-k/3) \qquad (25)$$

When these equations were used, the values found for N_{1B} and N_{1C} were 160 turns and 141.6 turns, respectively. Integers must be used, of course, so that N_{1C} was left at 142 turns. In fact, to improve the precision slightly, the positions of the taps feeding the C-unit were changed to 38 turns from center. In many cases, especially if the cores are matched, this refinement in the design would not be necessary.

Usually, for the master unit only of a transistor-magnetic inverter, provision must be made to insure positive self-starting. The resistor R_S in Fig. 11 serves this purpose and is quite conventional. The feedback windings had 26 turns on each side of center in each inverter, and base resistors R_B were 5.6 ohms, each.

Voltage waveforms from the self-lock-

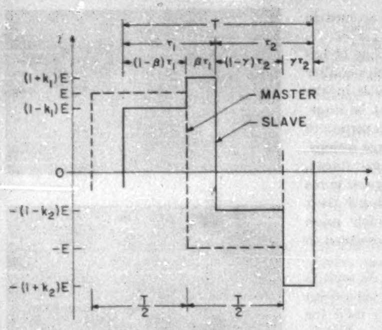


Fig. 13. Winding-voltage waveforms for asymmetric 2-phase inverter

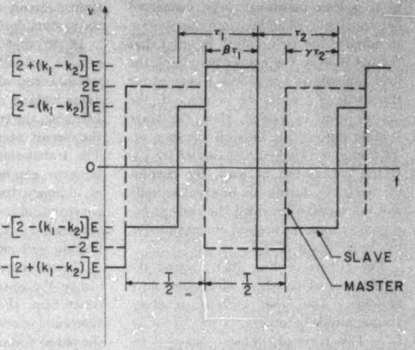


Fig. 14. Collector-to-collector voltage waveforms for asymmetric 2-phase inverter

ing inverter are shown in Fig. 12. Rectangular collector-to-collector voltages appear in Fig. 12(A), and the rectangular collector-to-collector voltage waveform for the master, together with a stepped collector-winding voltage waveform from each of the two slaves, in Fig. (12B).

Asymmetric Self-Locking

The theory leading to equation 10 can be generalized further to include the effect of feeding a slave inverter from taps on the master main winding, which are displaced from the center tap by unequal numbers of turns. This matter has academic interest, but in addition, since errors can be made in locating these taps, the results of the analysis have some practical importance in assessing the effect of such errors upon phase displacement. A self-locking polyphase inverter, under these conditions, can be called asymmetric.

Several other types of asymmetry suggest themselves but will not be analyzed here. For example, the master inverter can be made asymmetric by connecting the d-c supply to some tap other than a center tap on the collector winding. Similarly, in a slave, the number of turns on the two collector windings need not be equal. These three basic types of asymmetry and combinations thereof have been tested, and in all cases, self-locking action has been obtained. To indicate the type of operation which can be expected, analytical results and some experimental data are presented here for the type of asymmetry first mentioned.

It will be assumed that both master and slave are symmetrical in all respects except for location of the outlying taps on the master-inverter main winding, which location will be specified by calling their tap ratios k_1 and k_2.

The waveforms of the steady-state winding voltages are shown in Fig. 13. The linkage ratio ρ and the volt-time area of the master λ_A are given as before by equations 8 and 5, respectively, from which, and with use of the additional symbols defined in Fig. 13, equations can be written to describe the steady-rate constraints

$$\rho T/2=(1-\beta)\tau_1(1-k_1)+\beta\tau_1(1+k_1) \qquad (26)$$

$$\rho T/2=(1-\gamma)\tau_2(1-k_2)+\gamma\tau_2(1+k_2) \qquad (27)$$

$$T=\tau_1+\tau_2 \qquad (28)$$

$$\gamma\tau_2=T/2-(1-\beta)\tau_1 \qquad (29)$$

Equations 26 through 29 can be solved for β, γ, τ_1 and τ_2 in terms of the linkage ratio ρ, the tap ratios k_1 and k_2, and the period T of the master inverter. The results are

$$\beta=\frac{1}{2}\left[1+\frac{2(\rho-1)}{2k_1-\rho(k_1-k_2)}\right] \qquad (30)$$

$$\gamma=\frac{1}{2}\left[1+\frac{2(\rho-1)}{2k_2+\rho(k_1-k_2)}\right] \qquad (31)$$

$$\tau_1=\frac{T}{2}\left[1-\frac{(k_1-k_2)(\rho-1)}{k_1+k_2}\right] \qquad (32)$$

$$\tau_2=\frac{T}{2}\left[1+\frac{(k_1-k_2)(\rho-1)}{k_1+k_2}\right] \qquad (33)$$

Equations 30 and 31 reduce to equation 9 if $k_1=k_2=k$. Furthermore, in this case, $\tau_1=\tau_2=T/2$.

It is interesting to note also from equations 30 through 33 that if $\rho=1$, as in Card's inverter for example, β and γ are equal and each is one-half. That is, phase-lock angles different from 90 degrees cannot be obtained from Card's circuit merely by driving the slave from asymmetrical taps in the master main winding.

As an example of asymmetric self-locking, suppose $\rho=1.1$. For a symmetrical inverter, a tap ratio k of 0.3 would give 120-degree phase displacement which corresponds to $\alpha=2/3$. Suppose, however, that tap ratios k_1 and k_2 of 0.3 and 0.2 are used. The positive and negative portions of the slave period would have durations of 0.98 $T/2$ and 1.02 $T/2$, respectively. The deviation from $T/2$ is very small. The phase displacement is no longer uniquely defined, however, unless it is considered to be the phase displacement between the fundamental of the slave waveform and that of the master. The values of β and γ are 0.704 and 0.696, respectively. Had the tap ratio been 0.2, the value of α would have been 0.75, which corresponds to a phase displacement of 135 degrees. Roughly, therefore, the use of different tap ratios to feed a single slave produces a phase displacement in between those charac-

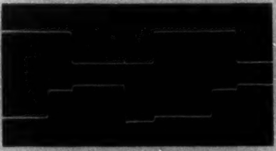


Fig. 15. Oscillograms of collector-to-collector voltages for asymmetric 2-phase inverter

Top—master
Bottom—slave

teristic of the two different tap ratios when used in two separate symmetrical 2-phase circuits.

Also, if an asymmetric circuit of this type is used, the waveform of the collector-to-collector voltage in the slave is not rectangular, but stepped as shown in Fig. 14. A typical oscillogram of the slave collector-to-collector voltage for tap ratios of 0.3 and 0.0 is shown in Fig. 15, together with the waveform of the master inverter collector-to-collector voltage.

Conclusions

Self-locking inverters can be designed for any phase-lock angle between 0 and 180 degrees. Slave output voltage can lag or lead the master output voltage by any desired angle within this range. Design tolerance on the phase-lock angle is very small. Several slave inverters can be driven by one master inverter.

The phase-lock angle is a function of the linkage ratio of the master and slave transformers and of the tap ratio which characterizes the position of the driving taps on the master transformer. Each of these parameters may be chosen from a wide range of values. Phase-locking transients occur without any oscillation around the steady-state phase-lock angle.

Collector-to-collector loading of the slave inverters may be preferable when a rectangular load-voltage waveform is required. However, loading of a slave inverter by means of a secondary load winding is entirely possible and, within fairly wide limits, has a loading effect upon the master inverter which is quite tolerable. Symmetrical self-locking circuits are generally preferable, but locking phenomena of the same sort occur when asymmetry is introduced.

References

1. PHASE LOCKING OF SWITCHING-TRANSISTOR CONVERTERS FOR POLYPHASE POWER SUPPLIES, A. G. Milnes. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, Nov. 1955. pp. 587–92.

2. TRANSISTOR-OSCILLATOR INDUCTION-MOTOR DRIVE, W. H. Card. *Ibid.*, vol. 77, Sept. 1958, pp. 531–35.

3. APPLICATIONS OF SATURABLE CORES IN TRANSISTOR POWER CONVERTERS, R. P. Putkovitch, T. M. Corry. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 14, 1958, pp. 874–83.

4. A MORE STABLE 3-PHASE TRANSISTOR-CORE POWER INVERTER, W. E. Jewett, P. L. Schmidt. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, Nov. 1959, pp. 686–91.

Magnetics in Doppler Signal Data Extraction

R. J. METZ
NONMEMBER AIEE

J. G. FAY
MEMBER AIEE

THE PURPOSE of this paper is to describe a solid-state approach to the Doppler radar data extraction problem using nonlinear magnetic techniques. The mechanization details will therefore be preceded by sufficient radar system background pertinent to this application.

Coherent radar systems are employed as an approach to using the Doppler shift phenomenon. The shift in frequency of the received signal spectrum is proportional to the radial velocity between the transmitting source and the target and, if suitably measurable, will become a source of information about a particular target parameter, range rate $\dot{R}$, which is radial velocity. This measurement is based upon the comparison of the received frequency with the originally radiated frequency, which requires that the radiation be coherent. To describe a target, the concept of a "target space" is useful. The volume of possible target parameters surveyed by the radar system is quantized along radial and angular co-ordinates. The angular quantization is a function of the radiating transducer (antenna) and will not be further considered. Further quantization in the radial co-ordinate takes place within the confines of an individual angular quantum. The target parameters along the radial co-ordinate which are of interest are range rate $\dot{R}$ and range R and may be termed target space within a beamwidth. In many air-borne systems it is necessary to time-share a single antenna system for transmission and reception which naturally leads to a pulsed system. This pulsed radiation of the r-f (radio-frequency) energy is also a suitable form of modulation for range measurement in which the range quanta are established by the time duration of the transmitted pulse. Thus, a pulsed coherent radar establishes a target space in which range is quantized and measurable and range rate is continuously measurable.

Detection of signals in random noise, such as is introduced by all practical antennas and receivers, is materially aided by reducing the bandwidth through which the corrupted signal [received signal plus noise $(S+N)$] passes before a detection decision is made. Detection constitutes a decision as to whether the signal is or is not present, i.e., a binary or 2-choice decision. Furthermore, the decision process is inherently statistical. Once the Doppler shift in frequency is measurable it must be measured to some degree of resolution. Frequency may be measured by using a large number of contiguous narrow-bandwidth channels covering the Doppler shift frequency range of significant information. In effect, the continuum of values of frequency is quantized such that the bandwidth of one of the frequency channels is 1 quantum. This frequency channelizing coincides with the desire for bandwidth reduction made necessary by the requirements of signal detection. As a result of this bandwidth reduction and the associated quantization of the Doppler spectrum, the most desirable target-space parameter to use as a basis for detection decisions is range rate. Therefore, by detecting Doppler shift rather than the pulse modulation envelope time delay, the system becomes a pulsed-Doppler-type coherent radar system. The range quanta then provide the resolution necessary to identify the physical volume in which a target is detected on the basis of observations of its range rate through the Doppler shift. This leads to the concept of a suitable detection and measurement system based on $\Delta R \times \dot{R}$ cells or "boxes." The ΔR parameter represents range increments in which target returns may appear and is obtained by time-gating the re-

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 14, 1960; made available for printing September 21, 1960.

R. J. METZ and J. G. FAY are with Westinghouse Electric Corporation, Baltimore, Md.

The authors acknowledge the important contributions made by N. L. Weinberg and B. W. Poehlman in the mechanization of this equipment and the help of Dr. T. J. Pula in writing this paper.

turned signals into a number of separate receiver channels as determined by resolving the interpulse period of the transmitted waveform into segments as dictated by the width of the transmitted waveform. The receiver system then becomes a multichannel system in which the $\dot{R}$ (Doppler frequency) resolution is carried out by a separate bank of contiguous frequency filters in each channel. To examine the output condition of this matrix of $\dot{R}\times\Delta R$ boxes, some method of high-speed sequential interrogation is required. An obvious method would be the use of a multipole commutator switch. However, the speed and contact synchronization requirements are beyond the realm of mechanical switching techniques. The body of this paper will be devoted to a discussion of the considerations involved in utilizing nonlinear magnetics in the mechanization of a static approach to this data extraction problem compatible with high-speed digital data processing techniques.

Functional Requirements

These considerations lead to functional requirements for the mechanization of equipment for Doppler radar signal data processing. This processing may be viewed as comprising two distinct parts: data extraction and data organization. The purpose of data extraction processing is to examine the radar receiving system spectral output and make meaningful decisions as to the presence or absence of targets.

The data organization process makes use of these detection decisions, which are binary or 2-state (YES–NO) in nature and, according to established rules, arranges this information into sets describing a particular target. This form of processing is handled by conventional digital data handling techniques and will not be discussed further. The data extraction process, using a bank of contiguous frequency filters as the primary measuring element, leads to the following set of functional requirements.

CHANNELIZATION OF DOPPLER SPECTRUM

To perform measurements of the continuum of Doppler frequencies in which the information resides, the available frequency spectrum is separated into discrete segments by means of filters having bandwidths commensurate with the desired $\dot{R}$, target radial velocity, resolution. The radar system characteristics, including target environment, determine the extent of the useful Doppler spectrum and maximum permissible resolution. These requirements determine the number of filters necessary per range element. The pass-band shape of the filter is also important in meeting the measurement requirement.

COHERENT INTEGRATION

The bandwidth of each channelizing filter is inherently related to the available observation or integration time preceding a detection decision; this is termed coherent integration.[1] The band pass of each filter also aids in the process of detection of signal in random noise by reduction of the noise spectrum. This is termed predetection filtering.

ENVELOPE DETECTION

All targets with the same range rates (corresponding to the same Doppler shift of the transmitted carrier) will occur as outputs of the channelizing filters and, when detected, will contain the necessary target existence information. The center frequency of the channel filter labels this shifted carrier frequency. The amplitude modulation of the shifted carrier frequency relates to target "strength," i.e., radar cross section and range. The envelope demodulation of this shifted carrier permits examination of the amplitude characteristics to make a detection decision as to the presence or absence of a target. Since this type of demodulation is instrumental in arriving at the detection decision, it is termed envelope detection.

POST-DETECTION INTEGRATION

In order to aid the detection decision process, the envelope demodulated output of the predetection channelizing filters can be subjected to further integration, often called post-detection integration. This will increase the probability of detection for a given signal power to average noise-power ratio (signal power related again to target strength).[2] The post-detection integrator should be an ideal integrator to avoid the adverse effects of unequal weighting in the time domain.[3] Such effects are demonstrated by integrators having low-frequency corners. The integrator will be periodically examined after an active integrating interval to determine the total integrated demodulated signal plus noise. This process is termed interrogation and should leave the integrator in a zero or cleared state after each examination. This requirement is necessary to achieve statistical independence from one integrating interval, i.e., "look-time" or observation period, to the next. Resorting to post-detection integration reduces the number of frequency channels required in covering a given spectral region by permitting the use of a larger quantum (greater filter bandwidth) of frequency if allowed by the resolution requirements on frequency ($\dot{R}$) measurements.

INTERROGATION

The first step in the actual data extraction process is interrogation. Interrogation, in effect, asks a question of any particular sequence of cascaded filter, envelope detection, and integration functions as to the existence of signal energy of sufficient magnitude to represent the existence of a target. This cascade of signal processing functions represents an $\dot{R}\times\Delta R$ box and is termed a frequency detection channel (FDC).

Interrogation of a particular FDC must make available a signal representing the integrated level of $S+N$ observed by the channel since the last interrogation. In addition, the integrating element must be returned to a normal or zero condition, following (or during) interrogation. The results of this interrogation must be suitably switched from many FDC's to a decision mechanizing function. Alternatively, the decision-making function may be mechanized within each FDC and the YES–NO decisions suitably switched to a digital data organizing entity. Interrogation must proceed at high speed so that sampling of all FDC's will not require time comparable to the magnitude of the system look period. Depending upon the exact nature of the data organizing process, the interrogation process may be required to be under continuous command control on a step-to-step basis.

Mechanization of Frequency-Detection Channel

INTEGRATION

It can be seen that the manner of mechanizing the interrogation and post-detection integration functions is the key to meeting the system functional requirements with minimum size and weight equipment. A square-loop magnetic core was chosen as the integrator in view of its easy compatibility with the switching requirements of the interrogation or sampling process. The relationship between mathematical integration and magnetic-core operation is immediately evident through Faraday's law. Not so obvious, however, is the fact that such an integrator represents the integrated output quantity as a magnetization state of the core. A useful way to characterize such a representation is that the output of the device is stored in the device; that is, it is not available as an electric signal but exists as a state which must be subjected to examination in some manner. Consequently, such an integrator might best be termed a storage integrator and as such does not provide an electric signal output. Interrogation is essentially a process of measuring this magnetization state and subjecting this measurement to a suitable decision criterion. The characteristics of square-loop magnetic materials lend themselves naturally to high-speed switching systems.[4] The object in using magnetic techniques as the backbone of the interrogation requirement is to have the same core which mechanizes a cross point in a switching matrix also act as an integrating element without interfering with the interrogation process. An integration mechanism compatible with interrogation requires that the core be time-shared between these two functions. To achieve this efficiently, both driving agencies operating upon the core must be decoupled in some manner. This means that Faraday-law integration utilizing a low-impedance, or voltage, source cannot be used in practice. The required decoupling is most effectively developed when both driving sources are high-impedance, or current, sources. This suits the interrogation requirement naturally but poses a problem from the point of view of integration. Clearly, dependence on Faraday's law in integral form is not possible if the input signal is a current rather than a voltage source.

$$\Delta\phi(\propto \Delta M_{\text{core}}) = \int e\,dt$$

where e is the independent (integrate-drive input signal) variable and $\Delta\phi$ is the magnetization state representative of the integral of the input signal.

Integration using a current source is an accumulation process in which an irreversible flux change (remagnetization) takes place whenever the core threshold is exceeded by a signal current. In this mode of operation, Faraday's law merely expresses a measure of the resulting remagnetization (proportional to $\Delta\phi$) rather than providing the mechanism.[5] It is important to note that such a mechanism can be postulated only for a limited range of integrate-drive current amplitudes. The use of a storage integrator core and the mode of interrogation postulated, i.e., a "destructive" read-out, meet the system requirement of statistical independence between successive observation periods. The act of interrogation returns the post-detection storage integrator (PDSI) core to its cleared state and ready for the next integration period. The integration mechanism requires a unidirectional drive magnetomotive force (mmf). This is obtained naturally by the envelope detection function which supplies signal information to the integrator element. Another requirement which must be met is that of decoupling, which stems from the essential nature of the core; namely, that it is always a transformer.

Envelope Detection

Envelope detection of the predetection filter output can be conveniently mechanized to aid the core integration process by using a grounded-base transistor as a coupling element between the filter and PDSI core. In so doing, the decoupling requirement imposed by the inherent transformer nature of the core is effectively met. This is important from the point of view of not interfering with the circuit action required for the interrogation process. In addition, energetic coupling between the predetection filter and a storage integrator core undergoing interrogation must be avoided because false alarms would then be generated by the interrogation process alone. This process is similar to asking a leading question in which the mere asking predetermines the answer.

The envelope detection process desired is essentially one of amplitude demodulation in which the output-signal frequency (limited to the pass band of each frequency channelizing filter) is the Doppler shifted carrier. The target information sought resides within the effective amplitude modulation of this pseudocarrier caused by range and size characteristics of the target and the admixture with white Gaussian noise falling within the spectral region defined by the predetection filter. The purpose of detection in this case is to give a measure of the amplitude of the incoming wave.

Envelope detection is usually considered to be a demodulation rather than detection process since, in detection, the central idea is to make a decision in favor of one of two possible alternatives. The purpose of envelope detection in this system organization is to provide a measure of the amplitude characteristics for subsequent threshold comparison. Therefore, one approach to envelope detection consists in subjecting each individual half-cycle of the incoming carrier wave to a threshold and remembering in an accumulative fashion the extent to which the instantaneous input waveform amplitude exceeds this threshold. This function is provided by the post-detection integrator. The successive exceedings of this threshold level are then stored in the core element until a subsequent interrogation time. The action of such a threshold within the envelope detection function more truly makes the process one of detection rather than merely demodulation.

It can now be seen that the use of a square-loop magnetic-core element combines the requirements of envelope detection (in this specialized context) with post-detection integration by an interrelated exploitation of inherent core properties. The practical achievement of this action depends upon the use of a suitable coupling element between the filter and the core. This is accomplished by a class A grounded-base transistor in which the input is a current signal from the filter into the emitter-base (E-B) diode. The output of this coupling is then a collector current α times the summed emitter current. The emitter input impedance approaches zero, an ideal condition for summing current sources. The exact value of this bias current from the point of view of maintaining class A operation is not critical and thus can be manipulated during manufacture as a convenient ad-

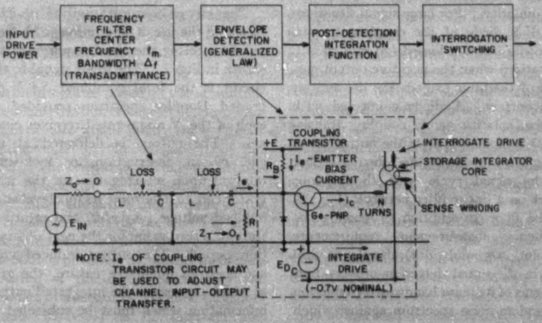

INPUT DRIVE POWER
FREQUENCY FILTER CENTER FREQUENCY f_m BANDWIDTH Δ_f (TRANSADMITTANCE)
ENVELOPE DETECTION (GENERALIZED LAW)
POST-DETECTION INTEGRATION FUNCTION
INTERROGATION SWITCHING
COUPLING TRANSISTOR
+E
R_B
I_e -EMITTER BIAS CURRENT
INTERROGATE DRIVE
STORAGE INTEGRATOR CORE
SENSE WINDING
$Z_o \rightarrow 0$
LOSS
LOSS
L
C
L
C
i_e
i_c
N TURNS
Ge-PNP
E_{IN}
R_I
$Z_T \rightarrow 0_r$
E_{DC}
INTEGRATE DRIVE
(-0.7V NOMINAL)
NOTE: I_e OF COUPLING TRANSISTOR CIRCUIT MAY BE USED TO ADJUST CHANNEL INPUT-OUTPUT TRANSFER.

Fig. 1. Functional circuit schematic of frequency detection channel organization

justment variable for compensating for variations in core threshold (Fig. 1). Such an adjustment also provides a simple change of input to output transmission characteristics of each FDC.

Frequency Channelizing Filter

The filters used for quantizing the spectral region of interest are organized into contiguous, or comb, arrangements termed a "filter bank" and driven by a range-gated receiver channel output. It should be noted that the output drive from the receiver may be characterized as having a constant average noise-power output as the result of a very tight delayed, i.e., referenced, automatic gain control system.

The circuit configuration selected for the filter was defined first by input–output circuit restrictions coupled with signal-power limitations and second, by the fractional bandwidth requirements as determined by signal spectra and resolution requirements. The first consideration required a voltage input-current output, that is, a transadmittance type of transfer function, as determined by the nature of the receiver output driving stage (since it must drive parallel filter elements) and by the low input impedance nature of the grounded-base coupling transistor. The Doppler spectral region of interest (24 kc) and the required 2-db (decibel) bandwidth of 400 cps (cycles per second) lead to fractional bandwidth requirements compatible with *L-C* (inductance capacitance) filter technology using powdered Mo-permalloy linear magnetic cores. The bandwidth requirements and realizable *Q*'s permitted a coupled series-resonant 2-mesh configuration in which coupling is from the input series-resonant mesh to output series-resonant mesh. The output termination ideally approaches a short circuit; the current into this termination is the output quantity. For large signal situations the E-B diode of the coupling transistor could be back-biased (go into class *C* operation) when the negative output current of the filter exceeds the d-c emitter bias current. A diode connected with inverse polarity across coupling transistor E-B diode, prevents the output mesh from being open-circuited during a part of the signal cycle with the obvious effect on a linear filter network (Fig. 1).

The filter, in addition to meeting the frequency measurement requirements (resolution ∝ bandwidth), also plays a vital role in the signal detection process, first in terms of its noise bandwidth (reduction of random noise spectrum against which a threshold signal must compete), which is predetection filtering, and second in terms of its coherent integration action for a time period established by the 3-db bandwidth (500 cps→2 ms (millisecond) coherent integration time).

Switching Requirements

The need for interrogation of the integrated output contained by the storage integrator core as a magnetization state leads to certain required circuit conditions. It has been shown that the need for time-sharing the integrator core between the function of integration and its subsequent interrogation is partly met by the use of current source drives and particularly by the use of the grounded-base coupling transistor for the integrate drive. The change of magnetization state back to a cleared state as a consequence of interrogation causes an induced voltage of a certain polarity to appear on all circuits coupled to the core.

This induced voltage is the desired flux information in time-derivative form and can be reconstituted by suitable circuit techniques as an electric signal proportional to this magnetization change. This induced voltage is sensed in a common output circuit (sense winding) for further processing without affecting other channels or being affected by other channels associated with the same filter bank.

The time-sharing of the integrator core is obtained by consideration of the relative amplitude of the interrogate-drive and integrate-drive current sources. Interrogation flux switching takes place in a very short period of time, of the order of several microseconds (μsec), while integration flux change (magnetization away from the cleared state) takes about 10,000 μsec. This leads to a relationship between the current drive amplitudes such that the interrogate drive, in effect, "overpowers" the integrate drive to achieve time-sharing.

The circuit organization of a FDC based on the use of square-loop core techniques meeting the cited requirements is shown in Fig. 1. The input signal to this circuit is the total combined $S+N$ degraded Doppler spectrum provided by each of the N range-gated receiver channels. The output is delivered at the end of an observation, or look-time, interval upon interrogation of the integrator core to determine its flux level as an induced voltage ($\propto d\phi/dt$) and return its state to a zero value for the next observation period. The output induced voltage is of an impulsive nature, the area of which represents the integrated output information which must be subjected to a threshold decision.

Data Extraction Commutation Considerations

A data extraction method based on the use of $\dot{R}\times\Delta R$ boxes (FDC) leads to certain commutation or time-sharing requirements.

First, the total number of decision (YES–NO) communication channels between the data extraction mechanization and the data processing entity charged with data organization should be minimized by time-sharing. The most logical approach would be to establish a decision channel for each range-gated receiver. This decision channel would then be time-shared by successive decisions made upon the $\dot{R}$ subdivided boxes within a given range-gated receiver channel. Whether these decisions within one ΔR channel would be made in ordered or random sequence is immaterial for the moment. Another logical possibility would be to interrogate $\Delta R\times\dot{R}$ boxes at random on command and transmit results through only one decision communication channel to the organizing entity.

As already stated, the interrogation process destructively changes the magnetization state of the integrator core to a cleared state. Through this action a time integral of voltage is induced in any and all circuits associated with the core. Thus, the nature of this process is essentially one of "ideal" mathematical differentiation: $e=d\phi/dt$. This differentiation operation converts the magnetization state of the integrator into an electric signal. Since this signal is in derivative form, it must be integrated to provide an electric signal representing the magnetization (flux) level of the integrator core just prior to interrogation. The detection decision must then be made on the basis of this reintegrated signal.

At this point a dichotomy of approach to the commutation problem presents itself. Shall the detection decision be mechanized separately for each FDC with the binary decision output signals combined by suitable binary logic circuits for subsequent transmission over one or several commutation channels? If so, the detection criteria and process must be mechanized many times. Or is it better to time-share the integrator core differentiated output signals with a single electric signal integrator and subject this reintegrated signal to a single mechanization of the detection criteria? The former method, when subjected to practical considerations of stability, "hardware" economy, and reliability, is not practical from the point of view of stability of mechanization of the detection criteria. This is particularly true if an integrator core approach is employed because the nature of the interrogation process makes electric reintegration necessary. The second approach, while commutating (analog) information to a single integrator-detection signal processing sequence, is simply mechanized when the integrator core is employed by series-connecting all sense windings of cores whose outputs are to be time-shared on a single circuit and taking advantage of the inherent properties of square-loop core materials to achieve the necessary noninteracting mixing type of function. This type of switching function is analogous to the OR binary logic functions required if the first commutation approach were followed. The square-loop core property of low initial permeability up to a certain threshold field (for subsequent domain-wall motion remagnetization switching) leads to a simple mechanization of this function by the simple expedient of threading the cores on a single wire.

The nature of a storage integrator, compared to what might be termed a signal integrator, needs further exposition in order to clarify certain actions carried out in this mechanization.

An example of electric signal integration is provided by a passive or active *R-C* (resistance capacitance) network integrator. The *R-C* type provides an output electric signal representation of the integral and usually develops this signal as a result of energy storage. On the other hand, a storage integrator based on a square-loop core with ideal remanence properties represents the integral as a state (level of remanent magnetization). In terms of information-bearing propagating electric energy, no signal is available from a storage, or state, integrator until some particular action is taken. Asking the core to reveal its magnetization state (interrogation) is accomplished by remagnetization to the cleared state. This action also accomplishes the integrator clearing requirement which is considerably simpler than the removal of stored energy required for resetting an electric signal integrator. This requirement for *R-C* integrators would impose a considerable burden in a many-channel system-mechanization because the interrogation would have to be achieved by a sampling action followed by the clearing action. Such a requirement seriously inhibits rapid commutation of many FDC's into a common decision channel.

The system by which the array of FDC's is interrogated is not restricted in any way by the integrator core concept. The sequence of scanning is determined only by the system of data organization employed by the data processing entity. This follows as a consequence of the inherent speed and flexibility of organization for which magnetic-core switching techniques are well known. Scanning may be achieved in either ordered or unordered sequence. The former is usually known as "sequential" scanning and does not require an addressing scheme since time is one address. The latter is known as random scanning and obviously requires some sort of address scheme for gaining access to a particular location in the integrator core array. A practical requirement in a radar system such as previously described is the sequential commutation of like $\dot{R}$ elements in $N-\Delta R$ channels into N decision channels. In effect, N detection decisions are made in parallel and dispatched to the data organizing function. The data organization method is very closely related to the data extraction system. In many cases, the nature of this relationship requires that scanning be under complete start–stop control of the organizer. These diverse requirements are readily met with magnetic-core switching techniques which are completely compatible with the interrogation switching required by the integrator core concept. The most obvious scanning system evolves as the solid-state static analog of an N pole commutator switch, capable, however, of virtually immediate start–stop operation. Clearly, such a requirement, if $N \rightarrow 20$, is beyond the realm of mechanical devices. This is especially true if speed considerations are such that the time available for interrogation is a relatively small part of that between successive interrogations, of the order of 1 ms or less. With magnetic switching techniques, interrogation speeds of 4–8 μsec per step are readily achieved with no problems of contact synchronization.

Post-detection Integration Based on Magnetics

Previous considerations have shown that an integrator based on the accumulative magnetization change of a square-loop magnetic core would satisfy the pertinent requirements in mechanization of an FDC. The central problem is the attainment of such an accumulating "integration" under the action of a current-source drive coupled from the predetection filter. The functional requirement of envelope detection is provided by the threshold nonlinearity of the core. Since only the presence or absence of so-called threshold signals ($S/N<1$) is of interest, this threshold action is not at all objectionable providing it is controllable and predictable. This action effectively thresholds the frequency channelizing filter output signal before integration. As a consequence, mathematical analysis becomes difficult, if not impossible, and an experimental approach to developing the required integration action is indicated.

Straightforward interpretation of these requirements in terms of the parameters used to describe square-loop magnetic core properties, allowed elimination of whole classes of commercially available materials and cores. The threshold and integration requirements dictated that the static hysteresis loop have very steep sides and a well-defined threshold H field for irreversible remagnetization (switching). Since the grounded-base coupling transistor has a current transfer (gain) approaching unity (0.98), the current available to act as an integrate drive for a storage integrator core is essentially the filter output, which is $i_{out}=Ge_{in}$. In the interests of drive-power economy, it is necessary that this current be reduced to a minimum. This forces the use of a toroidal geometry having the smallest possible length of magnetic path. Once the hysteresis loop shape requirement eliminates the use of manganese–magnesium ferrites, attention focuses solely on ultrathin metallic-tape magnetic materials. The use of stainless-steel bobbin-supported tapes of 1 mil and less thickness with material outside diameter of less than 100 mils, provides a feasible solution in terms of drive current. Such core configurations are available in production quantities at low prices.

The threshold action giving rise to envelope detection is such that the threshold should remain invariant throughout the time period for which the core is required to accumulate irreversible magnetization changes. In addition, such a core must be capable of slowly switching under the action of a constant-level drive mmf. Practical square-loop core materials present problems in the attainment of such a response. The usual situation in which a slow time rate of change of flux arises, is when a core is driven from a voltage source. Here, the time rate of flux change is constrained to a value related to the applied drive voltage by Faraday's law. In the present instance, current source drive causes a critical dependence of $d\phi/dt$ upon exact material parameters. From the point of view of maximizing the dynamic range of the interrogated flux-level read-outs, it is desirable to improve the signal-to-noise ratio of this read-out mechanism. (Note that this is not to be confused with the

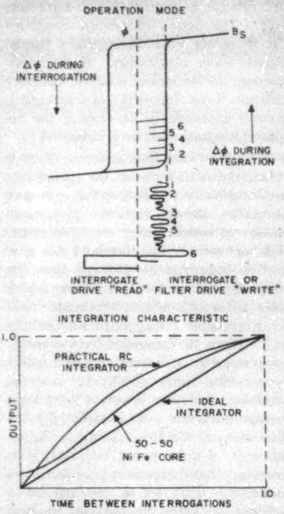


Fig. 2. Storage integrator core concept

S/N involved in the detection, and that the noise referred to in the read-out mechanism is systematic in nature and thus completely predictable. The noise in the radar detection case is fundamentally random and thus involves statistical treatment.) This signal to noise is fundamentally related to the squareness property of the core, B_R/B_S. Maximization of core squareness on a flux basis (time-integrated induced voltage) is accomplished by making full use of the available magnetization range of the core, positive to negative saturation remanence. To make full use of this available flux change with a current-source drive it is necessary that virtually all intermediate remanent states from negative to positive saturation remanence be characterized by essentially the same magnitude of threshold field for irreversible flux change. Were this not so, to achieve the full available flux with a constant-current drive magnitude would require such a magnitude as to cause the core to remagnetize very rapidly and thus defeat the purpose—to integrate slowly over a relatively long time (about 10 ms).

Thus, the achievement of integration as well as the functional envelope detection depend critically on the threshold properties. The core material selected must optimize this property. Of the commercially available square-loop metallics, 50-50 Ni-Fe (nickel-iron) alloys approach these requirements to a usable degree. The experimental verification of this threshold can be achieved by observing the static hysteresis loop. Unfortunately, determination of a truly static hysteresis loop for very small samples of material is an extremely difficult task. Some form of pulse test mechanization action is a much more practical approach since magnetization state changes are most easily measured by electric observation of the corresponding coupled flux variation followed by subsequent electric signal integration. However, this is only true when the electric signal integration is performed for short times.

The evaluation method which evolved during the material and core-selection phase was essentially a pulse test in which actual system operating conditions were simulated to some degree (Fig. 2). Because of testing philosophy, system functional requirements, and commercial availability of small elements, the core chosen was an 86-mil groove diameter bobbin core of 1/2-mil 50-50 Ni-Fe subjected to optimum annealing conditions by the manufacturer to develop performance characteristics in a reproducible, simulated test. The nature of this test was to apply a current source drive of such a magnitude, in relation to core threshold, to require a specified time to cause the full available flux change to be incurred; a period at least as long as the desired system observation or integration time. In effect, the test revealed that a particular integrator core actively observes the output of the predetection filter for at least 10 ms. This property is identified indirectly by observing the integration (accumulation) performance. This is done by interrogating the core with a short, high-amplitude current-drive pulse to read destructively its magnetization state into suitable sensing and measuring circuits. The interrogation is conducted periodically at a rate inversely proportional to the desired integration time.

Such a measurement leads to a so-called integration characteristic in which the integrated output (i.e., the core flux change accumulated from a cleared state during integration and subsequently read out during interrogation) characteristic is taken under constant integrate drive conditions and shows the normalized magnetization change accumulated during integration as a function of normalized time available for integration. An ideal integrator on such a basis with one-half the time would show one-half the output. Properly processed cores of 50-50 Ni-Fe alloys can be made economically to show properties useful in a detection system requiring integration. The exact nature of simulation of the integrate drive is very important since the integration mechanism has almost a first-order dependence upon it. This is to be intuitively expected since a *reductio ad absurdum* would prompt one to make the bias current as large as the core threshold field and, consequently, only a very small signal current would be required to cause a flux change. Obviously, such action is physically unattainable. As a practical matter, values of d-c bias current of the same order of magnitude as the signal-drive current will afford satisfactory integration performance.

A simple model for the integration mechanism may be visualized by considering the statics and dynamics of domain walls. It is essentially the properties of the domain wall microstructure which control the physics of the remagnetization process exploited in this application for achieving an accumulative type of integration. For simplicity, consider an idealized single crystal of ferromagnetic material having a geometry leading to a 2-domain configuration. The two domains have net magnetization vectors oppositely directed (180-degree wall). Experimental measurements confirm the theory that the velocity with which the wall moves is proportional to the effective H field.[6] The motion of the domain wall: perpendicular to the magnetization direction, represents the growth of one domain at the expense of the other, i.e., the remagnetization process. Consider a sensing area at right angles to the plane of the wall velocity vector. The rate of area change is obviously the rate of flux change observed by a sense circuit coupled to this area. Integrating the wall velocity once with respect to time gives the displacement while a similar operation to the voltage induced in the sensing circuit ($e = d\phi/dt$) gives the total flux change observed. Consequently, the displacement of the wall is measured by the time integral of the induced voltage. An important point to keep in mind is that this displacement process is irreversible to the extent that upon completion it is represented by a static-state condition of magnetization. Consequently, the wall displacement is proportional to the time integral of the effective H field, to which the velocity was directly related. The effective H field is the applied field minus the threshold H field for domain wall motion (H_o). Ideally this H_o parameter is a constant for all possible magnetization states of the material, which is certainly the case for the simple model

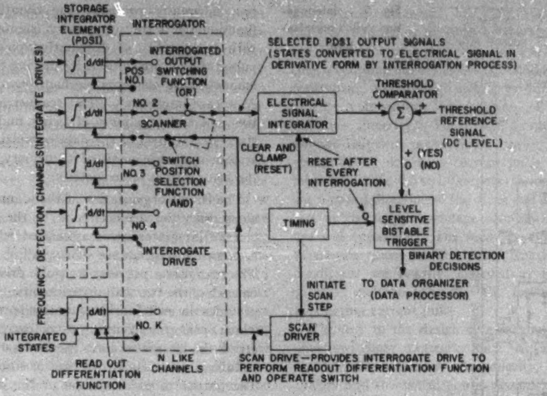

STORAGE INTEGRATOR ELEMENTS (PDSI)
INTERROGATOR
FREQUENCY DETECTION CHANNELS (INTEGRATE DRIVES)
INTERROGATED OUTPUT SWITCHING FUNCTION (OR)
POS NO. 1
NO. 2
SCANNER
NO. 3
SWITCH POSITION SELECTION FUNCTION (AND)
NO. 4
INTERROGATE DRIVES
NO. K
INTEGRATED STATES
READ OUT DIFFERENTIATION FUNCTION
N LIKE CHANNELS
SELECTED PDSI OUTPUT SIGNALS (STATES CONVERTED TO ELECTRICAL SIGNAL IN DERIVATIVE FORM BY INTERROGATION PROCESS)
ELECTRICAL SIGNAL INTEGRATOR
THRESHOLD COMPARATOR
THRESHOLD REFERENCE SIGNAL (DC LEVEL)
CLEAR AND CLAMP (RESET)
RESET AFTER EVERY INTERROGATION
+ (YES)
0 (NO)
TIMING
LEVEL SENSITIVE BISTABLE TRIGGER
BINARY DETECTION DECISIONS
INITIATE SCAN STEP
TO DATA ORGANIZER (DATA PROCESSOR)
SCAN DRIVER
SCAN DRIVE—PROVIDES INTERROGATE DRIVE TO PERFORM READOUT DIFFERENTIATION FUNCTION AND OPERATE SWITCH

Fig. 3. Decision channel functional organization based on storage (state) integrator concept

chosen. For the case of a polycrystalline material of thin tape in a spiral-wrap toroidal geometry, such a simple approach would not apply. However, this simple model is of considerable conceptual value when employing the empirical test procedure already discussed.

Commutation Circuit Mechanization

Decision Channel

The decision channel, Fig. 3, comprises a signal processing cascade of the interrogated core output switching function, an electric signal integrator, a threshold comparator, and binary signal standardizing and timing functions. The latter three functions are grouped together and considered as a threshold detector. The first function is mechanized by a common series-connected sense winding through all the PDSI cores associated with one ΔR detection channel.

The commutated integrator core outputs appear as unipolarity voltage signals on the sense winding. This may be a 1-turn or several-turn winding depending upon the characteristics of the electric integration circuit which operates on these signals. The voltage signals are of relatively small magnitude, about 100-mv (megavolt) peak, with volt-second "areas" (flux) up to several hundred mv-μsec but, fortunately, are developed from equivalent voltage generators with extremely low source impedances measured in milliohms. To avoid extraneous pickup systematic noise problems the geometry of the sense winding must be carefully controlled.

Consistent with the distributed inductance of the sense winding, with the low initial-permeability integrator cores as lumped L contributors, the sense circuit should be terminated in as low a resistance load (input impedance of electric integrator) as possible. This is in order to maximize the signal power delivered from the interrogated storage integrator core and consequently reduce the power-gain requirements in the threshold detector circuit. The lower limit on this terminating resistance is that the bandwidth of the sensing circuit transfer function (voltage in, current out) must not degrade the induced signal waveshape since the signal processing sequence of differentiation-integration would then be invalidated. This bandwidth limitation is most readily expressed in terms of the L-R (inductance-reactance) time constant. For relatively slow, say 5-μsec, switching of the integrator core the signal degradation experienced for an L_{sense}/R_{term} of 0.1 μsec (approximately 1-mc (megacycle) bandwidth) is negligible. For a 60-FDC system the inductance of 1-turn sense structure of reasonable geometry can easily be kept below 1 μh (microhenry). For multiple turns the lumped L contribution of each integrator core goes up by a factor of N^2, thus necessitating an increase in R termination to meet bandwidth requirements. Note that the output signal power is $\propto e^2$, where e is the voltage induced by the integrator core switching, which goes up as N.

The threshold detector circuit organization is shown in Fig. 4. The linear amplifier and Miller-type integrator G blocks are each mechanized by three cascaded grounded-emitter transistor stages with over-all loop negative feedback gain stabilization. The linear amplifier function preceding electric integration was made necessary by the requirement of integrator clearing and clamping of the signal integrator, accomplished by operating on the virtual ground summing node of the operational type of integrator. Consequently, pedestal effects and gating transients required that the input signal level to the operational integrator be at least several volts. The bandwidth requirement on this linear amplifier for the aforementioned core switching time of approximately 5 μsec is established by considerations similar to those posed for the sense-winding mechanization of the interrogated output switching function. The linear amplifier G block must be essentially direct-coupled up to the integrator resetting and clamping function. This is a direct consequence of the fact that total volt-second integral observed at the output of the sense circuit as a time function depends directly upon the nature of the information patterns being processed. However, the action of interrogation (once every observation period) of all associated storage integrator cores causes this integral to zero at least once

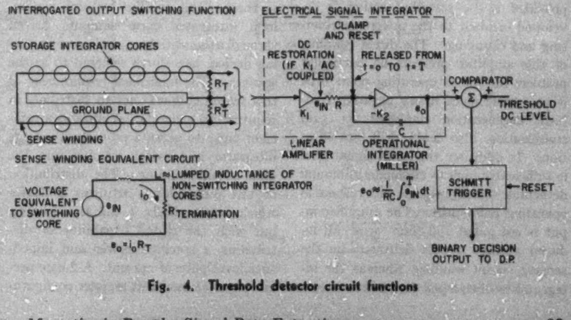

INTERROGATED OUTPUT SWITCHING FUNCTION
STORAGE INTEGRATOR CORES
R_T
GROUND PLANE
R_T
SENSE WINDING
SENSE WINDING EQUIVALENT CIRCUIT
L≈LUMPED INDUCTANCE OF NON-SWITCHING INTEGRATOR CORES
VOLTAGE EQUIVALENT TO SWITCHING CORE
e_{IN}
i_o
R TERMINATION
$e_o = i_o R_T$
ELECTRICAL SIGNAL INTEGRATOR
CLAMP AND RESET
DC RESTORATION (IF K_1 AC COUPLED)
RELEASED FROM t=0 TO t=T
K_1
e_{IN}
R
$-K_2$
e_o
C
LINEAR AMPLIFIER
OPERATIONAL INTEGRATOR (MILLER)
$e_o \approx \frac{1}{RC}\int_0^T e_{IN}\,dt$
COMPARATOR
THRESHOLD DC LEVEL
SCHMITT TRIGGER
RESET
BINARY DECISION OUTPUT TO D.P.

Fig. 4. Threshold detector circuit functions

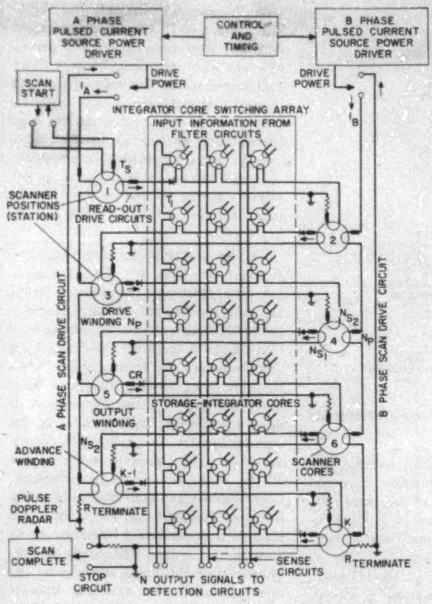


Fig. 5. Interrogation switch organization

every look-time interval. For observation intervals of the order of 10 ms, a low-frequency corner requirement for the linear amplifier *G* block of somewhat less than 1 cps would be necessitated if significant level shifts are to be avoided during a look interval. Achievement of video amplifiers with such low-frequency corners with transistors is difficult; consequently, a direct-coupled approach is more feasible. In addition, direct coupling avoids blocking problems due to input and output circuit transients. Since the gating required by the integrator acts in the linear amplifier output circuit, this circuit may be a-c capacitor coupled provided a d-c restoring action is developed incident to the integrator resetting and clamping. The direct coupling of this amplifier presents no particular problem insofar as operating point drift is concerned since heavy negative feedback degeneration employed for gain stabilization also adds drift stabilization. In addition, the input signal level is sufficiently high to eliminate minimum detectable signal problems with respect to operating point drift. The amplifier input is not gated and thus "sees" all induced voltage signals delivered by the sensing circuit winding, whereas the integrator is unclamped just before a core interrogation read-out signal voltage is delivered and integrates only up to the maximum possible duration of any such signal (by which time the threshold comparison has been made) after which the integrator is cleared and clamped.

Interrogation Switching

As previously noted, a most useful requirement for the interrogation of the integrator core array is in an ordered sequence. Consequently, the *N* integrator cores associated with the same $\dot{R}$ value are simultaneously interrogated by a current drive of sufficient power capability. To achieve this an interrogation switch is mechanized jointly by the square-loop integrator core elements driven through a sequential magnetic-core switching device, henceforth referred to as the scanner. The functional requirement of this switch is that it act to distribute in sequence to a number of stations a sufficient pulse to actuate properly a storage integrator and switch its flux into the sense-circuit winding. The distribution of this pulse clearly requires a device which is essentially a unitary counter, but with the added capability of distributing a specified power and impedance level pulse of current. A 2-core-per-bit magnetic-core shift register configuration inherently combines the counting function (as a ring or unitary counter) with a capability of distributing power pulses of a nature suitable for driving the magnetic-core elements mechanizing the *N* storage integrators. This scanner has the inherent capability of being under start–stop control of the data organizing system as a consequence of its memory of the last position interrogated.

The circuit organization of the scanner along with the drive circuits to the *N* storage integrator cores associated with each switch position are shown in Fig. 5. The scan driver provides a-c pulse drives on each of the two scan-drive circuits. It provides the switch-drive actuation power which performs counting and current drives the integrator cores (performs distributing) associated with that position. The output of each position of this sequential switching network is the interrogate drive for the associated integrator core elements of the switch by means of series-connected interrogate drive windings. The integrator core elements are inherently a part of the interrogation switching due to their square-loop characteristics and the noninteracting mixing switching function which this mechanizes.

While the scanner bears an organizational similarity to a 2-core-per-bit magnetic-core shift register, certain unique application requirements in this case cause a somewhat different design approach. The first is the need for achieving efficient power transfer to a given switch output as an interrogate drive-power pulse with good current waveform characteristics, such as amplitude, duration, and transition times. The usual *S*/*N* requirements and capabilities of such devices charged with transmission of binary information are not sufficient in this case. The small pulse outputs on "unselected" positions during an associated scan drive would be detrimental to the successful commutation of the integrated flux levels of the particular interrogated integrator core. While the threshold nonlinearity of the integrator core aids this requirement to some degree, additional steps must be taken to insure that the dynamic range of the integrator core-sense winding output circuit not be decreased by any other imperfection but the inherent noise due to nonunity squareness of the integrator core alone. From the point of view of economical generation of the scan drive-power current pulse, note that only one core on a scan-drive circuit is switched at any particular time out of say 50, unselected positions. (Note that a switch such as this for 100 positions would have 50 cores associated with each scan-drive circuit.) These goals were met by resort to 50-50 Ni-Fe 1/2-mil tape stainless-steel bobbin cores to reduce core noise flux due to nonsquareness and to reduce air-core reversible inductance to a minimum. This minimization of reversible inductance of a scan-drive circuit permits a reasonable mechanization of the scan driver which could easily become very inefficient and expensive. To take care of the remaining noise flux of a scanner core coupled into an interrogate drive circuit, the necessary shift-register coupling diode was purposely made one with poor forward transient recovery characteristics. In a ring counter application such as this, slow reverse transient recovery characteristics are not particularly detrimental since adequate time for carrier recombination will always transpire even in a large, fast sequential switch before the need to block reverse voltage develops.

The forward transient recovery characteristic found in relatively high-conductance silicon junction diodes of high reverse breakdown voltage rating was found adequate to absorb effectively the noise volt-seconds of a nonswitching (unselected) scanner switch core. Consequently, the undesired corresponding noise current in an interrogate drive circuit was reduced to a negligible value for the application. This property is usually considered a detriment in these devices and is largely uncontrolled on the part of the manufacturer. Considerable source and unit selection was necessary to develop an adequate supply for the systems built. However, the attainment of this characteristic should not be too difficult for large quantities. All that is involved is making a diode with a relatively thick base region to develop a delay time for minority carrier injection and transport from the heavily doped side of the junction of the ohmic contact on the base side. Such a device would present no real problem in manufacture as the usual requirement is as thin a base region as possible from the point of view of high-frequency performance. The effect observed is essentially equivalent to an inductance, $\int e\,dt = \Delta i L$, where for a given volt-second integral a large value of L serves to reduce the current change. However, this equivalent inductance is purely dissipative in character and consequently does not present the inevitable stored-energy release problem. Unfortunately this type of diode action does lead to greater volt-second absorption in the interrogate drive circuit thus necessitating a somewhat greater flux capacity for the scanner core. However, since some excess volt-second absorbing capability is required in addition to the storage integrator core elements, the use of such a coupling diode reduces the need for resistance "swamping," usually a much greater volt-second loss because of a resistor's linearity. The diode forward voltage transient property can easily be examined by driving a current pulse of large amplitude (0.5 to 1.0 ampere) through the forward direction of the diode and observing the junction voltage drop during the rise transition. The recovery characteristics under natural recombination conditions are readily observed as the junction voltage following the forward conducting current pulse.

The key to the design of a magnetic-core switching network such as the scanner lies in how economically, in dollars and incidental power dissipation, the necessary current-source drive pulse can be developed. As previously described, a contributing factor in an application of this nature would be the scan-drive line inductance, i.e., linear, reversible, energy storage-type inductance (not the core switching volt-seconds). Since each scanner core switch position output drives a variable load (the special coupling diode and its transient characteristics) as well as 0 to N storage integrator cores switching from some partial remanent state to the cleared remanent state, the drive power must be developed by a current source. This is required to achieve uniform switching response and standard cleared remanent magnetization states from the storage integrator cores in spite of varying information patterns in the FDC's. The scanner design for this application in terms of storage integrator core load, interrogate drive-pulse width, and transition times (related to ultimate speed of interrogation) required a 1-ampere driver with a 40-volt back-voltage capability.

By making use of high-gain bandwidth product silicon mesa transistors (*2N697*), negative current feedback can be effectively applied to the generation of moderate speed (0.5-μsec transition time) current pulses in the ampere range. To eliminate "ringing" of the high collector junction parasitic capacitance of the power transistor when a highly inductive load of nonswitched, shuttled, scanner cores with several turns is driven, a rather novel circuit approach has been developed for this application; see Fig. 6. The problem is avoided by recognizing that the ringing is due to a high-impedance driving source acting on an underdamped L-C circuit (L of cores, C of collector junction) during a relatively fast pulse-transition time. By arranging the output stage as a nonfeedback emitter follower during the drive transition times, the L-C circuit "sees" a low-impedance voltage source drive and is well damped and consequently free of oscillatory disturbances. When the current builds up to a certain value, as determined by a precision current-sensing R in the output circuit developing a voltage signal for error comparison to a reference, the feedback loop is closed and the output stage, though emitter loaded, appears as a current source drive to the output circuit for disturbances, such as switching cores, having frequency spectra commensurate with the loop bandwidth. The current reference

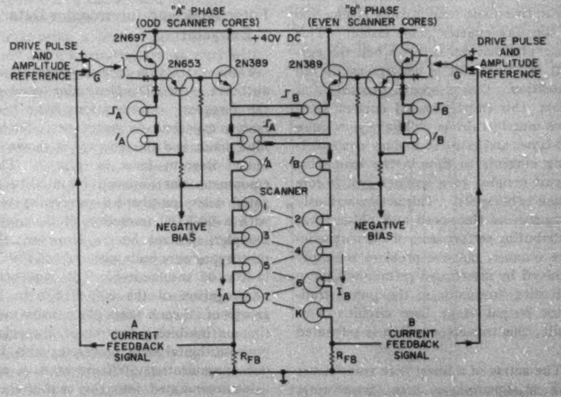


Fig. 6. Functional schematic of scan driver and associated scanner circuits

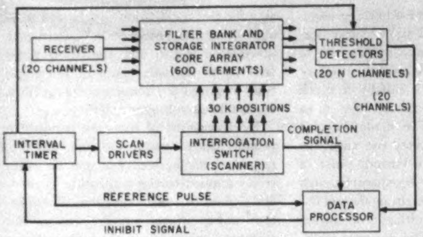


Fig. 7 (above). **Filter bank and interrogator block diagram**

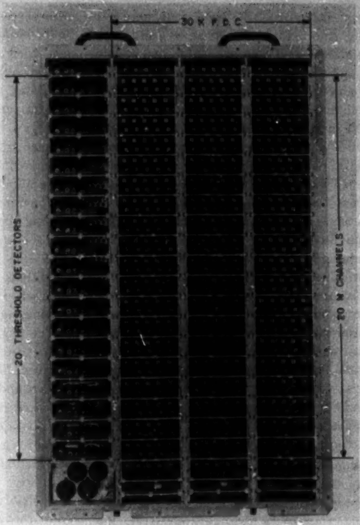


Fig. 8 (right). **Filter bank and interrogator unit**

is provided in this mechanization by a Zener diode. The error comparison and transition from open-loop voltage drive to closed-loop feedback amplifier current drive is made by a fast recovery silicon diode sensing the difference between the current feedback voltage signal and the voltage reference of the Zener diode.

This basic concept, however, does require a simple alteration to function effectively with a scan-drive-circuit load. It must be remembered that the scan-drive circuit is a reactive load in both a linear and a nonlinear sense. The linear reactive characteristic is represented by the reversible magnetic-energy storage parameter—inductance. The nonlinear reactive nature stems from the fact that when a switch core on, say, the odd numbered scan-drive circuit is reset, the next core on the even numbered scan-drive circuit is set, i.e., the count state is advanced. Realizing that set and reset flux changes mean both polarities of induced voltage, it can be seen that a voltage is induced into the off, or not driven, scan-drive circuit which adds to the d-c output stage power-supply voltage. For an emitter-loaded n-p-n transistor, this would tend to cause conduction in the base-emitter diode circuit. This same phenomenon occurs when the current in the circuit is undergoing its fall transition. Unless special precaution is taken, this emitter-loaded current feedback amplifier driver will have slow pulse fall times and cause secondary demagnetizing currents to flow in the scan-drive circuit coupled to a scanner switch core which is being set. This action seriously compromises the speed and pulse-power distributing performance of the magnetic-core scanner, but the problem is simply resolved by inserting a germanium p-n-p switching transistor in the power transistor output stage base circuit. Normally, this transistor switch is saturated ON.

The action of a linear core transformer and a square-loop core transformer coupling the E-B circuit of this switch to the output scan-drive circuit current is to "open" the base-drive circuit and permit a pull-down resistor to return the output base to a negative (for an n-p-n stage) bias voltage to keep its E-B diode back-biased. This action occurs for the fall of the drive pulse through the linear transformer coupling and for the set-induced voltage pulse by means of the cross-coupled square-loop core pulse transformers. This modification permits fall-time performance of the same level as the rise time, about 0.5 μsec minimum.

Filter Bank and Interrogator Data Extractor

Data extraction equipments for an air-borne pulsed Doppler radar based on the foregoing considerations have been built as experimental prototypes. Such a filter bank and interrogator is shown in block diagram form in Fig. 7. This equipment was designed for a 20-range gated detection channel receiver system with a 60-FDC resolution of the useful Doppler spectral band. However, the prototypes were built with only 30 FDC's of the 60 mechanized. The sequential interrogation of the 600 FDC's in 30 groups of 20 each takes place under continuous inhibiting control of a special-purpose digital data processing unit for data organization. Whenever data are being assimilated into the central data processor as a result of data organizing operations, the interrogator action is stopped until such time as the data processor is ready to receive more target detection decision information.

The performance of such a mechanization of data extraction is measured in terms of false-alarm rates and probability of detection versus input S/N ratio (sine waves additively combined with Gaussian noise in the receiver input circuits). Evaluation of the performance of square-loop magnetic cores as post-detection integrators has been considered elsewhere.[2] These measurements and the conclusions drawn from them relate P_D versus S/N ratio at the input to a particular frequency detection channel for certain false-alarm rates for a single 30-FDC range-gated detection channel as mechanized in the 20-channel prototypes.

The mechanical features of the packaging are illustrated in Fig. 8. A primary consideration in the packaging centered around the organization of the wiring layout of the storage-integrator core array and geometrical factors relating to the sense winding. The package was designed to minimize the sense-circuit length and area by arranging the storage-integrator core elements closely against a wiring-terminal slotted ground plane. This served to reduce stray pickup and maximize sense-circuit bandwidth as was previously discussed. This approach caused the interrogate drive lines to increase in length which represents an increased volt-second loss due to current-resistance drop and can easily be circumvented by care in design. As a result of this packaging approach, form follows function.

The prime requirement of air-borne equipment is minimum size and weight for performing a given function. In addition to inherent advantages in the commutation mechanization by using magnetic switching techniques rather than mechanical devices, the core approach allows considerable size reduction in the linear filter mechanization by virtue of the low-impedance design approach. It also allows smaller linear inductors—by far the greatest contributor to size and weight in a narrow-band linear filter. In addition, for a 1,200-FDC system, the number of active and passive components per FDC must obviously be held to a minimum. The magnetic storage-integrator core approach does just this by the multifunction per component mechanization which it permits.

References

1. An Analysis of Coherent Integration and Its Application to Signal Detection, K. S. Miller, R. I. Bernstein. *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y. vol. IT-3, Dec. 1957, pp. 237–48.

2. Rapid Detection of Coherent Radar Signals in Noise, R. J. Metz, J M Walker, N L. Weinberg. *Convention Record*, Institute of Radio Engineers, pt. 8, Mar. 1960.

3. A Statistical Theory of Target Detection by Pulsed Radar, J. I. Marcum. *Report no. RM 754*, Rand Corporation, Santa Monica, Calif., Dec. 1954.

4. Magnetics for Computers, A Survey of the Art in 1958, J. A. Rajchman. "Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication*, Aug. 1958, pp. 179–221.

5. Self-Saturating Magnetic Amplifiers (book), G. E. Lynn, T. J. Pula, J. F. Ringelman, F. G. Timmel. McGraw-Hill Book Company, Inc., New York, N. Y., 1960, pp. 81–98.

6. The Influence of Pulsed Magnetic Fields on the Reversal of Magnetization in Square-Loop Metallic Tapes, D. S. Rodbell, C. P. Bean. "Conference on Magnetism and Magnetic Materials," *AIEE Special Publication T-78*, Oct. 1955, pp. 219–27.

Control of Frequency and Phase Displacement in Transistor Converter Circuits by Means of R-C Networks

F. V. KADRI
ASSOCIATE MEMBER AIEE

A NEW TYPE of adjustable-frequency transistor converter circuit, described in Part I of this paper, differs from conventional transistor converters and from transistor converter circuits incorporating frequency adjustment discussed in previous papers.[1–3] In the new type, the operating frequency is independent of the applied direct voltage—not proportional to it. The switching rate is determined by the time constant of an *R-C* (resistance-capacitance) network so that no saturating element is required. By making either the resistor or capacitor variable, the frequency is easily controlled over wide limits, and once a setting is made, no appreciable change in frequency occurs despite large variations in the input voltage.

Described in Part II is a transistorized regulated power supply, operating from a d-c source and containing an *R-C* network for control of the phase displacement between a pair of separately generated square wave voltages. Thus, the rms, or average value of the sum of the square waves at the output, can be maintained constant.

Part I

This division of the paper deals with a transistor converter circuit, utilizing an *R-C* network for providing wide-range frequency adjustment, frequency being independent of the applied voltage.

The operation of d-c to a-c static converters, using a pair of switching transistors and a saturating transformer with a rectangular-hysteresis-loop core, is amply covered in the literature.[4–6] In these circuits, the operating frequency is determined by both the applied direct voltage and the saturating transformer. Novel methods have been devised for permitting the frequency of transistor converters to be varied readily. One early method employs control windings on a twin-core arrangement, by which means the total flux swing can be adjusted. More recently, schemes have been developed which transfer the switching function from the converter transformer to a small auxiliary saturating element, thereby furnishing a means of varying the frequency over broad limits. In contrast to these is a transistor converter herein described in which the switching rate is determined by an *R-C* network; consequently, the dependence of frequency upon input voltage is eliminated. This is accomplished without sacrificing the inherent self-oscillatory feature of the converter while, in addition, a simple means of varying the frequency over wide limits is made available.

CIRCUIT DESCRIPTION

An adjustable-frequency transistor converter employing an *R-C* switching network is shown schematically in Fig. 1. The center-tapped primary winding is designated as N_{p1}, N_{p2}, and the base-drive windings for switching transistors *Q*1 and *Q*2 are designated N_{b1} and N_{b2}, respectively. R_s is a starting resistor and V_{co} is the initial voltage on the capacitor. Winding N_a is an additional winding on the transformer core. The voltage developed across it, as well as the base voltages, alternate each half-cycle and together generate a current which charges the capacitor *C* through the variable resistor R_a in alternating directions each half-cycle. Although a common-emitter

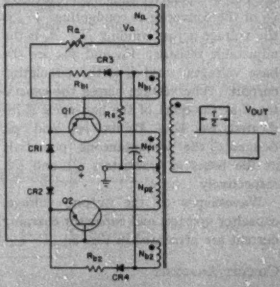


Fig. 1. Schematic diagram of adjustable-frequency converter employing R-C switching

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted August 18, 1960; made available for printing September 21, 1960.

F. V. Kadri is with Bell Telephone Laboratories Inc., Whippany, N. J.

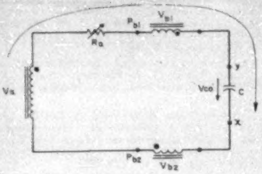


Fig. 2. Equivalent circuit of adjustable-frequency converter

circuit employing p-n-p transistors is given, any of the conventional circuits employing p-n-p's or n-p-n's can be used, provided the diodes and energizing winding are properly oriented.

A diode, *CR*2, is placed between the positive input terminal and the emitter of *Q*2 in order to eliminate a low-resistance path, which would otherwise exist across the capacitor by virtue of the base resistance R_{b2} and the emitter-base junction of *Q*1. Diode *CR*1 performs a similar function during the succeeding half-cycle. Also, during the latter part of the nonconducting half-cycle of each transistor, the winding end of its base resistor becomes negative relative to the positive supply terminal. Consequently, diodes *CR*3 and *CR*4 are inserted to prevent reverse current flow through the base resistors during these intervals. This prevents the occurrence of additional low-resistance paths around the capacitor. The current through R_t is small and can be neglected.

Because of the restrictions imposed by the four diodes, the capacitor charging current is confined to the loop composed of resistor R_a, capacitor *C*, and windings N_a, N_{b1}, and N_{b2}. This network determines the operating frequency of the converter. An equivalent circuit, made up of the elements of this loop, is shown in Fig. 2. The dots on the windings mark the positive polarity ends when transistor Q_1 of the converter is conducting. V_a is the voltage across winding N_a; R_a is the adjustable resistor; V_{b1} and V_{b2} are the base voltages; and i_c is the circulating current. The voltage across the capacitor *C* at the start of the half-cycle is indicated as V_{co}. Points p_{b1} and p_{b2} designate the instantaneous potentials at the bases of transistors *Q*1 and *Q*2, respectively.

Waveshapes of the output voltage, capacitor voltage, and capacitor charging current are presented in Fig. 3.

CIRCUIT ANALYSIS

A relationship between operating frequency and circuit parameters can be obtained by first deriving an expression relating the base potentials of the transistors to the circulating current. The positive (cutoff) potential, existing on the base of *Q*2 while *Q*1 conducts, gradually diminishes during the half-cycle because of the exponential decay of the current through R_a as the capacitor becomes charged. By applying Kirchhoff's voltage law to the equivalent circuit of Fig. 2, the equation for the current is found to be

$$i_c = \frac{E}{R_a} e^{-t/R_aC} \tag{1}$$

where

$$E = V_a + V_{b1} + V_{b2} + V_{co} \tag{2}$$

As seen in the equivalent circuit, the potential, p_{b2}, at the base of *Q*2 relative to the potential p_{b1}, at the base of *Q*1, is

$$p_{b2} = p_{b1} + i_c R_a - V_a \tag{3}$$

The voltage at the base of *Q*1 is held constant at a potential slightly negative with respect to the emitter while *Q*1 conducts. When the base voltage p_{b2} reduces to this same value, *Q*2 will conduct. Therefore, by setting $p_{b2} = p_{b1}$ in equation 3, substituting in the value of i_c from equation 1, and calling the time for the half-cycle $T/2$, the period *T* is determined. Its value is

$$T = 2R_aC \ln \frac{E}{V_a} \tag{4}$$

V_{co}, the initial voltage across the capacitor (contained in *E*), is easily determined. Because of complete circuit symmetry with respect to alternate half-cycles, the value of V_{co} must be equal and opposite in successive half-cycles. Furthermore, the voltage existing on the capacitor at the end of one half-cycle becomes the initial voltage in the next. Therefore, by determining the voltage remaining at the end of the positive half-cycle, the value of the V_{co} term in *E* can be obtained. This is readily done, since at time $t = T/2$, $p_{b2} = p_{b1}$, and the voltage on the capacitor, measured from *x* to *y*, is

$$V_{x-y} = V_{b1} + V_{b2} \tag{5}$$

Hence, the V_{co} term in *E* is also $V_{b1} + V_{b2}$ since it has been depicted as pointing from *y* to *x* in the equivalent circuit diagram.

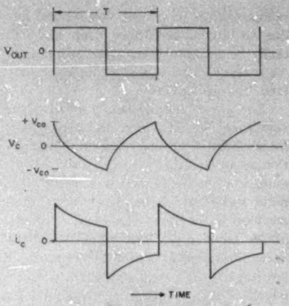


Fig. 3. Waveshapes of output voltage, capacitor voltage, and capacitor charging current in circuit of Fig. 1

Substituting this value of V_{co} into equation 4 and letting $V_{b1} = V_{b2} = V_b$, since the base voltages can be assumed equal, gives for the period

$$T = 2R_aC \ln \frac{V_a + 4V_b}{V_a} \tag{6}$$

Voltages V_a and V_b can be expressed in terms of the input voltage by using the transformer turns-ratios, therefore, the equation for the operating frequency, as a function of the circuit constants, is

$$f = \frac{1}{2R_aC \ln \dfrac{N_a + 4N_b}{N_a}} \tag{7}$$

where N_a and N_b are, respectively, the turns on winding N_a and the number of base winding turns.

If

$$N_a = 4N_b/(e-1) \approx 2.33 N_b$$

$$f = \frac{1}{2R_aC} \tag{8}$$

Because the described circuit depends upon the values of R_a and *C* for the maintenance of frequency, temperature compensation must normally be used for wide operating temperature ranges. This may take the form of a thermistor shunt across R_a.

Table I. Regulation of Output Frequency with Variation of D-C Input Voltage for Various Frequency Settings

	Percentage Change From Various Settings, Cps					
Input, Volts D-C	800	1,000	2,000	5,000	10,000	15,000
20 (−20%)	+0.8	0.7	+0.4	+0.2	+0.1	+0.1
25 (nominal)	0	0	0	0	0	0
30 (+20%)	−0.6	−0.2	−07	−0.2	0	−0.07

Fig. 4. Oscillograms of output waveshapes for various frequency settings

To insure that core saturation does not cause switching, the product of the turns on half the primary and the effective cross-sectional area of the core must fulfill the inequality

$$N_pA_c > \frac{V_{in\,max}\times 10^8}{4f_{min}\times B_{sat}} \tag{9}$$

where

N_p = the number of turns on half the primary

A_c = the effective core area in cm^2

B_{sat} = the saturation flux density of the core in gausses

$V_{in\,max}$ = the maximum anticipated input voltage in volts d-c

f_{min} = the minimum desired operating frequency in cps

A model designed to operate over a large frequency range with an input voltage variation of from 20 to 30 volts d-c was built and tested in the laboratory. It supplies a 40-watt load at 120 volts rms. Data taken on the unit at a number of frequency settings from 800 cps to 15,000 cps are presented in Table I, and oscillograms of the output waveforms for several of the frequency settings are pictured in Fig. 4.

Part II

Power regulation by means of *R-C* phase control of a pair of transistor converters is dealt with in the second division of this paper.

A variety of transistor regulator circuits have been developed for obtaining constant a-c and d-c outputs from a variable d-c source. Some of these schemes, although somewhat complex, are inherently efficient, while others, though simpler in construction, are poor in efficiency. Virtually all of these methods, however, may be categorized into one of two groups, depending upon whether regulation takes place prior to or after the power converter stage. Included in the former group are, for example, the case of a d-c regulator placed ahead of a transistor converter or the pulse-width modulation of a drive signal to a power stage; and in the latter group, those schemes whereby a power converter is permitted to run freely and control is applied at its output as, for instance, by a magnetic amplifier.

The method of obtaining regulated outputs, subsequently described, is both simple and efficient, but does not fall into either of the two categories mentioned. Regulation is effected directly at the power stage by combining the square wave output of a pair of transistor converters. The square waves are displaced in phase relative to one another by a controlled amount so that the rms, or average, value of their sum, is maintained constant, even though the magnitudes of the individual voltages vary in direct proportion to the supply voltage. An *R-C* network provides the necessary phase displacement as determined by a feedback signal.

For d-c or average a-c regulation, the feedback signal can be taken directly from the output. In the case of rms regulation, a d-c signal, proportional to the rms value of the output waveform, must be obtained. One way of accomplishing this is discussed in the Appendix.

CIRCUIT DESCRIPTION

A block diagram of the phase-controlled regulated power supply is shown in Fig. 5 to indicate, in a general way, how regulated a-c and d-c outputs may be obtained from an unregulated d-c source by generating a pair of square waves and controlling their phase relationship.

The particular circuit to be described is shown schematically in Fig. 6. At the top in this figure is an independent, or free-running, transistor converter composed of transistors *Q*1 and *Q*2, and transformer *T*1, and below it is a phase-dependent, or follower converter, consisting of transistors *Q*3 and *Q*4, and transformer *T*2. A phase control circuit—made up of the resistance R_a (fixed resistor R_f in shunt with transistor *Q*5), capacitor *C*, and winding N_a—interconnects the two converters. Windings N_{p1}-N_{p2} and N_{p3}-N_{p4} are bifilar primaries; N_{b1}, N_{b2}, N_{b3}, and N_{b4} are base drive windings; and R_{b1}, R_{b2}, R_{b3}, and R_{b4} are the base resistors. The feedback network is pictured in block diagram form only, since a number of arrangements may be used. One arrangement is shown in Fig. 7.

Each of the converters contains a secondary winding, N_1 and N_2. These give rise to voltages *V*1 and *V*2 for the upper and lower converters, respectively. The windings are combined in such a manner than when both converters are generating their positive half-cycles (transistors *Q*1 and *Q*3 conducting), the two voltages reinforce at the output, although when they are out of phase they oppose. A starting circuit is included in the upper converter, and consists of resistor R_s and diode *CRs*.

By means of winding N_a, wound on the core of transformer *T*1, a small signal voltage, V_a, in phase with the square wave of the independent converter, is made available as a phase-sensing source for the follower converter. This voltage is used to activate the *R-C* network, which is electrically coupled into the base circuit of the follower stage. Resistance R_a is variable because of the variable emitter-to-collector impedance of the transistor.

In Fig. 8, a simplified diagram of the phase-dependent converter circuit is shown, and included in it is the phase-control network with which it is coupled to transformer *T*1 of the independent converter. If it is assumed that the independent converter is generating a positive half-cycle and the follower is generating a negative half-cycle, the in-

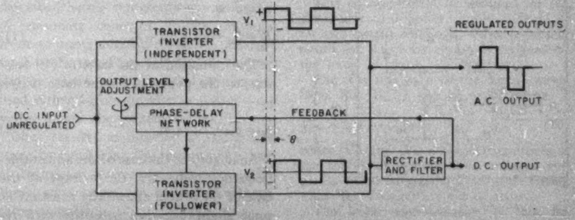


Fig. 5. Block diagram of power supply, regulated by phase relationship control of pair of transistor converters, using R-C network

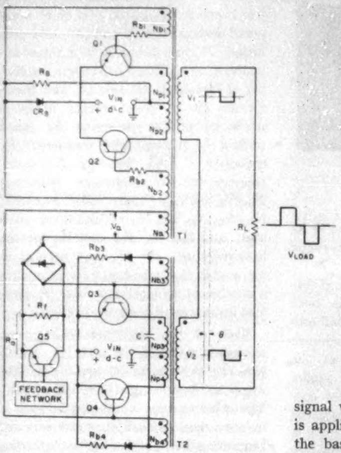
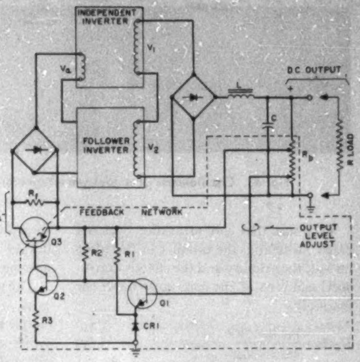


Fig. 6 (left). Schematic diagram of phase-controlled regulated power supply

Fig. 7 (right). Feedback circuit for d-c outputs in regulated power supply



stantaneous polarities of all windings will be as indicated, with the dots placed at the positive ends. This equivalent circuit applies for the period commencing at the instant the independent converter starts its positive half-cycle and ends just prior to the time the follower converter switches positive. Expressed in radians or degrees, this interval represents the angular phase displacement between the two square waves, and during this time the instantaneous secondary voltages V_1 and V_2 of Fig. 7 subtract at the output, their sum being zero if the magnitudes of V_1 and V_2 are equal, not actually a necessary condition for regulation to occur.

CIRCUIT ANALYSIS

In the circuit of Fig. 8, in which transistor Q4 is shown conducting, the signal winding N_a from transformer T1 is applied with its positive terminal at the base of Q4 and its negative terminal connected to the resistance R_a. Also, in this figure, the base-drive voltages are indicated as V_{b3} and V_{b4}, and the instantaneous potentials at the bases of Q3 and Q4 are designated as p_{b3} and p_{b4}, respectively. An exponentially decaying current, i_c, circulates around the outside loop in consequence of the signal voltage V_a and the two base voltages. V_{co} is the initial capacitor voltage.

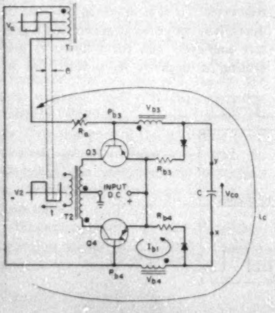


Fig. 8. Schematic diagram of follower converter with coupling network

During the period of phase delay, the potential p_{b3}, at the base of Q3, is positive relative to the potential p_{b4} at the base of Q4, its value being determined by the current i_c. When this base potential becomes sufficiently low, Q3 will conduct and the converter will switch. A relationship between the phase difference and the circuit parameters can be obtained if the manner in which i_c decays with time is first determined. This can be done in the same manner as for the adjustable-frequency converter of Part I.

The resulting expression for i_c is

$$i_c = \frac{E'}{R_a} e^{-t/R_aC} \tag{10}$$

where

$$E' = V_a + V_{b3} + V_{b4} + V_{co} \tag{11}$$

The potential at the base of Q3, relative to the potential at the base of Q4, is

$$p_{b3} = p_{b4} - V_a + i_c R_a \tag{12}$$

Analogous to the case of the adjustable-frequency converter, the voltage at the base of Q4 is held essentially constant at some slight negative potential with respect to the emitter while Q4 conducts. The length of time required for the base of Q3 to reach the same potential, measured from the instant the independent converter goes positive, is the time lag of the follower converter. It can be determined from equation 12 by setting $p_{b3} = p_{b4}$, putting in the value of i_c from equation 10, and solving for the time t. This is done herewith, t being called τ, the time delay.

$$\tau = R_a C \ln \frac{E'}{V_a} \tag{13}$$

The time delay normally ranges from a minimum of $0.01T$ to a maximum of $0.11T$ ($\approx$ 5-degree to 40-degree phase shift), where T is the period. Fig. 9 shows that for values of $\tau \leq T/4$ (90 degrees), the initial capacitance voltage, V_{co}, in equation 11 is zero. This is so because of the

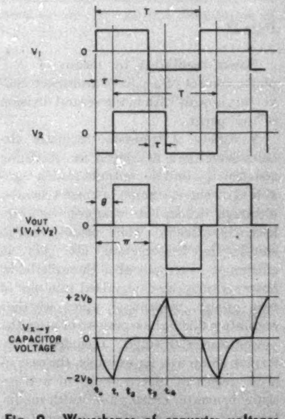


Fig. 9. Waveshapes of converter voltages, output voltage, and voltage across capacitor C for circuit of Fig. 7

approximately equal charge and discharge times during either half-cycle.

Since the a-c or d-c value of the output voltage waveform depends upon the amplitude of the voltage and the proportion of τ to T, or the angular phase shift, a determination of this quantity will provide a relationship between the output voltage and the circuit parameters.

The period of the independent converter depends upon the material and area of the core, the number of primary turns, and the applied direct voltage, so that, in accordance with the basic transformer equation

$$T=\frac{4N_iB_{i\,\mathrm{sat}}A_i}{V_{d-c}\times 10^8}\text{ seconds} \tag{14}$$

where

N_i = the number of turns on half the primary winding

B_i = the saturation flux density of the core in gausses

A_i = the area of the core in square centimeters

V_{d-c} = the direct voltage applied across half the primary winding in volts (transistor drop neglected)

$T=\frac{1}{f}$ = the period

Subscript i denotes that these values apply to the independent converter only

The period of the follower converter is also T, as indicated in Fig. 9, because:

1. The follower converter is designed so that, in the absence of external control, its operating frequency is significantly lower than that of the independent converter. Consequently, it has a longer natural period and cannot switch prior to time $T/2$.
2. Circuit symmetry causes the follower converter to lag the independent converter by equal amounts in commencing the positive and negative half-cycles; therefore, both operate at the same frequency.

The phase angle, designated θ in Fig. 9, is

$$\theta=\frac{\tau}{T}\times 2\pi\text{ radians} \tag{15}$$

Substitution of equations 13 and 14 into 15 gives the phase angle in terms of the circuit constants. The value of the phase angle is then

$$\theta=\frac{R_aC\ln\dfrac{E'}{V_a}\times 2\pi}{4N_iB_{i\,\mathrm{sat}}\,A_i/V_{d-c}\times 10^8} \tag{16}$$

N_i and A_i are constants, as is also $\ln E'/V_a$ (because both E' and V_a vary directly with V_{d-c}). If it is assumed that $B_{i\,\mathrm{sat}}$ is also constant, then

$$\theta=R_aCV_{d-c}\,K\text{ radians} \tag{17}$$

where K is a constant.

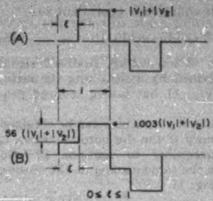


Fig. 10. (A) Output waveshape, and (B) wave developed to have an average value equal to rms value of output waveshape

This shows that the phase difference between the two square waves is directly proportional to R_a, C, and V_{d-c}. Because R_a is variable, the magnitude of θ can be adjusted to compensate for variations in the input voltage and thereby maintain the output constant.

Resistance-capacitance couplings can also be used to phase-lock two or more converters so that multiphase outputs can be obtained. Frequency regulation can be obtained if R-C switching is used in the independent converter.

Conclusions

Part I: By eliminating transformer saturation as the frequency determining factor in a transistor converter circuit, and employing instead an R-C switching network, these advantages are realized:

1. The operating frequency is constant, independent of the direct input voltage over a large input voltage range.
2. The frequency can be varied readily over broad limits.
3. Ordinary core materials can be used since no saturating element is required.
4. Switching losses are low because of the rapid switching afforded by the R-C network.
5. The absence of a saturating core considerably reduces switching spikes.

Part II: Phase-control regulation as described herein represents a novel way of obtaining constant voltage outputs from an unregulated d-c source. In contrast to the usual type of regulated power supply, where special control circuits are used either before or after a power converter stage, the present method accomplishes regulation directly at the power stage by combining the outputs of a pair of transistor converters. Consequently, these advantages are obtained without sacrificing size:

1. Inherently good efficiency since regulation is accomplished without the use of dissipative elements such as a series transistor or resistor in a series regulator.

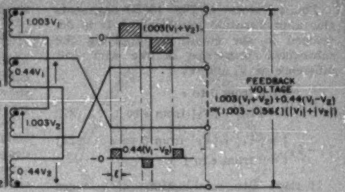


Fig. 11. The rms sensing circuit for phase-controlled regulated power supply

2. Circuit simplicity because of the direct method of obtaining regulation.
3. Reliable transistor operation because the full-square-wave drive signals insure adequate cutoff potential during the nonconducting half-cycle.
4. Reduced transformer insulation resulting from voltage-sharing of the secondary windings.
5. Lower current rating requirements for the transistors because of input current-sharing.

Appendix. The RMS Sensing Circuit

To obtain rms regulation in the phase-controlled regulated power supply discussed in Part II of this paper, a d-c signal, which reflects rms changes, must first be developed.

A method of obtaining a d-c feedback voltage, proportional to rms value of a "chopped" square wave, has been developed by W. Leonard and R. A. Stuart and has been put to practical use in a power supply employing magnetic amplifier regulation.[7] It is shown that the true rms value of the output, which is given by the equation

$$V_{\mathrm{rms}}=\sqrt{1-\epsilon}\;Vo \tag{18}$$

(where Vo is the magnitude of the output square wave and ϵ is the fractional part of a half-cycle during which the voltage is zero) can be approximated, to a high degree of accuracy, for values of ϵ up to 0.4, by the expression

$$V_{\mathrm{rms}}\approx(1.003-0.56\epsilon)\,Vo \tag{19}$$

This equation must be modified for the phase-controlled power supply because of the two output windings. An rms sensing circuit for the power supply is designed on the basis of these considerations:

The true rms value of the output waveform, Fig. 10(A), can be expressed as

$$V_{\mathrm{rms}}=\sqrt{1-\epsilon}\,(|V_1|+|V_2|) \tag{20}$$

where $|V_1|$ and $|V_2|$ are the magnitudes of the individual square-wave voltages and ϵ is the fractional part of a half-cycle during which the voltages are out of phase.

The linear approximation of this, for values of $\epsilon\leq 0.4$, is

$$V_{\mathrm{rms}}\approx(1.003-0.56\epsilon)(|V_1|+|V_2|) \tag{21}$$

The waveform of Fig. 10(B) represents this approximation graphically. It is developed in the circuit by applying the subsequent properties of the sum and difference of V_1 and V_2

$$V_1+V_2\begin{cases}=0 \text{ from } 0 \text{ to } \epsilon\\ =|V_1|+|V_2| \text{ from } \epsilon \text{ to } 1\end{cases} \qquad (22)$$

$$V_1-V_2\begin{cases}=|V_1|+|V_2| \text{ from } 0 \text{ to } \epsilon \qquad (23)\\ =0 \text{ from } \epsilon \text{ to } 1\end{cases}$$

If these two combinations of voltages are multiplied, respectively, by 1.003 and 0.44, the resulting equation is

$$1.003(V_1+V_2)+0.44(V_1-V_2) \qquad (24)$$

but (V_1-V_2) has a value of $|V_1|+|V_2|$ from 0 to ϵ and (V_1+V_2) exists only from ϵ to 1, when its value is also $|V_1|+|V_2|$, so that this equation is equivalent to

$$(1.003-0.56\epsilon)(|V_1|+|V_2|) \qquad (25)$$

as desired. A low-voltage feedback signal can be obtained by maintaining the ratios shown in Fig. 11 but taking reduced proportions.

Both rms and d-c outputs can be obtained simultaneously if the d-c output section is designed according to equation 21 and arranged in the same manner as the sensing circuit of Fig. 11.

References

1. A Variable Frequency Magnetic-Coupled Multivibrator, R. L. Van Allen. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, July 1955, pp. 356–61.

2. Application of Saturable Cores in Transistor Power Converters, R. P. Putkovick, T. M. Corry. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 14, 1958, pp. 874–83.

3. Recent Developments on Magnetic-Coupled Multivibrators, William A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, May 1960, pp. 106–12.

4. A Switching Transistor D-C to A-C Converter Having an Output Frequency Proportional to the D-C Input Voltage, G. H. Royer. *Ibid.*, vol. 74, July 1955, pp. 322–26.

5. Junction Transistors Used as Switches, R. L. Bright. *Ibid.*, Mar., pp. 111–21.

6. A New Self-Excited Square-Wave Transistor Power Oscillator, G. C. Uchrin, W. O. Taylor. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 43, 1955, p. 99.

7. A Magnetic-Amplifier—Transistor Power Supply for Missile Application, B. Mokrytzki, R. A. Stuart. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, 1959 (Jan. 1960 section), pp. 835–43.

Relaxation Oscillations at Multiple Frequencies of Even Subharmonics

NOAH S. PRYWES
ASSOCIATE MEMBER AIEE

THIS PAPER OFFERS a method of analysis for a parametric amplifier circuit. The circuit, composed of nonlinear inductance for parametric excitation, is designed to operate at one-half of the pump frequency. The Parametron[1] has proved useful as a general-purpose element for digital computers, having the potential capability of logic, storage, and amplification. It is both reliable and relatively inexpensive for computers of various sizes.

The analysis given herein provides physical understanding of the operation, in addition to quantitative design procedure for establishing:

1. Mode of energy transfer, output power, and gain.
2. Parameters of circuit components involved.
3. Voltages and currents in the circuit.
4. Impulse energy and timing required to switch the parametric amplifier on or off.

An analysis, based on assuming sinusoidal waveforms and quasi-linearity, is not sufficiently accurate to obtain quantitative design information. Instead, characteristics of nonlinear components are approximated by a number of linear segments and the relaxation oscillations are analyzed by first assuming that a large disturbance from the "off" mode is introduced. Then, the operation following this disturbance is studied, leading to the description of an "on" mode. In the discussion, a loss less (zero load) circuit is considered first. The manner of power transfer is then outlined and a practical case of nonzero load is considered. This leads to defining conditions of balance for input and output powers and stability.

Circuit and Characteristic

The circuit is shown in Fig. 1. The energy supply is in the form of a square wave current of i_{s0} amplitude, flowing in the supply circuit windings. The supply circuit current (i_s) is shown in Fig. 2. The direct current in the bias circuit biases the magnetic characteristics of the two cores. The magnetic characteristic of each core is simplified by approximating the unsaturated parts with vertical lines and the flux saturated parts with horizontal lines as seen in Fig. 3(A). Fig. 3(B) gives the flux (ϕ) versus gate circuit current (i_G) of one core and Fig. 3(C) of the other core. Adding the fluxes of the two cores, through series connection of the windings, the total flux (ϕ) versus the current i_G is obtained as seen in Fig. 3(D). Current saturation is achieved along parts *BC* and *DE* of the characteristic.

Operation is limited to the *BCDE* region of Fig. 3(D). A current saturable characteristic is thus obtained, as opposed to the flux saturable characteristic of Fig. 3(A). The characteristics for $i_s=+i_{s0}$ and $i_s=i_{s0}$ are drawn in Fig. 3(D), where

$$i_0=i_B\frac{N_B}{N_C};\ \Delta i_G=i_{s0}\frac{N_S}{N_G} \qquad (1)$$

During the period through which i_s stays constant, i.e., $i_s=i_{s0}$ or $i_s=-i_{0}$, only the gate circuit need be considered. The inductance is then a nonlinear element with the two flux-current characteristics of Fig. 3(D) corresponding to $i_s=i_{s0}$ and $i_s=-i_{s0}$. The parametric excitation expresses itself in the two changes, from one of these characteristics to the other, per supply frequency cycle.

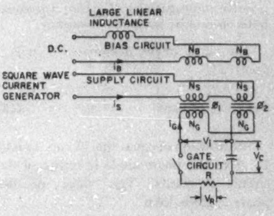


Fig. 1. Parametric oscillator circuit

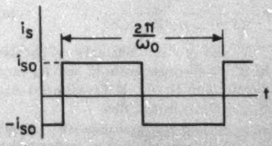


Fig. 2. Supply circuit current

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted June 23, 1960; made available for printing September 29, 1960.

Noah S. Prywes is with the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa.

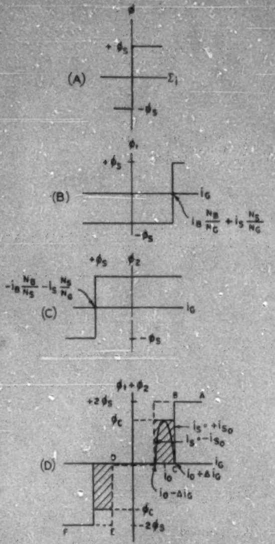


Fig. 3. Simplified characteristic of nonlinear inductance

The analysis is concerned with the case of a square-wave current supply rather than a sinusoidal supply. In addition to simplifying the analysis, the former allows for larger power output and hence larger gain, as will be clarified in more detail later.

The assumption of approximating the magnetic characteristics with horizontal and vertical segments neglects two important things:

1. The hysteresis losses which will be accounted for by an equivalent resistance, which is considered to be part of the resistance in the output circuits.
2. The flux change in the flux saturated portion of the magnetic characteristic. The simplifying assumption that there is zero change of flux as represented by horizontal segments in Fig. 3(A) introduces an error in the calculation of the power transmitted. However, a suitable correction can be made as will be shown.

Oscillations in Circuit Having Zero Resistance

When the gate circuit in Fig. 1 has the switch open, the capacitor c is charged until the voltage across it $V_c=V_{c0}$. If the switch in the gate circuit is closed at time $t=0$ when $i_s=i_{s0}$, the flux $\phi=0$ prior to the closing. At the instant of closing the switch, current will start flowing in the gate circuit, discharging the capacitor. From the characteristic in Fig. 3(D), the current i_G is limited as seen in equation 2. At

$$i_s=i_{s0} \qquad |i_G|\leq i_0+\Delta i_G \tag{2}$$

and at

$$i_s=-i_{s0} \qquad |i_G|\leq i_0-\Delta i_G$$

Initially, the resistance R is assumed to be zero, $R=0$. The circuit equations are

$$V_c=-N_G\frac{d\phi}{dt}$$

$$V_c=V_{c0_t}+\frac{1}{c}\int_0^t i_G dt \tag{3}$$

$$N_G\phi=\int_0^t V_c dt$$

The ϕ versus V_c relationship can be determined through integration and substitution of equations 2 and 3, resulting in

$$V_c^2=V_{c0}^2-\frac{2N_G}{c}(i_0+\Delta i_G)|\phi| \text{ for } i_s=i_{s0}$$

$$V_c^2=V_{c0}^2-\frac{2N_G}{c}(i_0-\Delta i_G)\,|\phi| \text{ for } i_s=-i_{s0} \tag{4}$$

These two relationships describe two families of closed trajectories, Γ_1 at $i_s=+i_{s0}$ and Γ_2 at $i_s=-i_{s0}$ in the ϕ versus V_c plane, as shown in Fig. 4.

At

$$V_c=\pm V_{c0},\ \phi=0$$

at

$$V_c=0,\ \phi=\phi_{01}>0$$

or

$$\phi=\phi_{02}<0$$

Therefore, equation 4 can be written in the form

$$V_c^2=\frac{2N_G}{c}(i_0+\Delta i_G)(\phi_{01}-\phi) \qquad \text{for } \phi<0 \text{ and } i_s=i_{s0}$$

$$V_c^2=\frac{2N_G}{c}(i_0+\Delta i_G)(\phi-\phi_{02}) \qquad \text{for } \phi<0 \text{ and } i_s=i_{s0}$$

$$V_c^2=\frac{2N_G}{c}(i_0+\Delta i_G)(\phi_{01}-\phi) \qquad \text{for } \phi>0 \text{ and } i_s=i_{s0}$$

$$V_c^2=\frac{2N_G}{c}(i_0-\Delta i_G)(\phi-\phi_{02} \qquad \text{for } \phi<0 \text{ and } i_s=-i_{s0} \tag{5}$$

From equations 3 and 4, or 5, the cycle period of the gate circuit oscillations is determined to be

$$T=\frac{4V_{c0}c}{i_0+\Delta i_G} \text{ for } i_s=+i_{s0} \tag{6}$$

$$T=\frac{4V_{c0}c}{i_0-\Delta i_G} \text{ for } i_s=i_{s0}$$

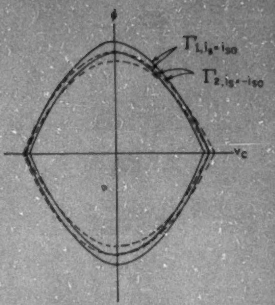


Fig. 4. Trajectories of the oscillations for $I_s=I_{s0}$ and $I_s=-I_{s0}$

and

$$\frac{T(i_s=+i_{s0})}{T(i_s=-i_{s0})}=\frac{V_{c0}(i_s=+i_{s0})}{V_{c0}(i_s=-i_{s0})}\left(1-\frac{2\Delta i_G}{i_0+\Delta i_G}\right)$$

$$=\sqrt{\frac{\phi_{01 \text{ or } 2}(i_s=+i_{s0})}{\phi_{01 \text{ or } 2}(i_s=-i_{s0})}}\left(1-\frac{2\Delta i_G}{i_0+\Delta i_G}\right) \tag{7}$$

Oscillations in Circuit with Resistor Losses

In Fig. 1, the load is connected in series with the capacitance and nonlinear inductance. (Similar conclusions are obtained if the load is connected in shunt with the capacitance.)

The circuit equation 3 becomes

$$V_c+Ri_G+N_G\frac{d\phi}{dt}=0 \tag{8}$$

Integrating equation 8, the trajectories in the $\phi-V_c$ plane, for the case of losses, $R\neq0$, in the gate circuit, are determined. There trajectories are for $i_s=+i_{s0}$

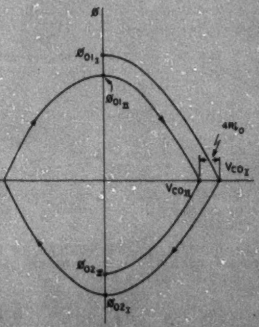


Fig. 5. Trajectory of operation at $I_s=\pm I_{s0}$ with $R\neq0$

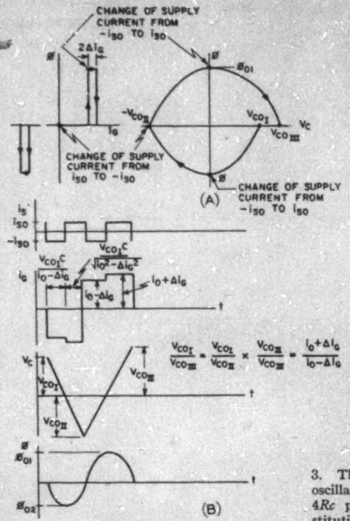

CHANGE OF SUPPLY CURRENT FROM $-i_{s0}$ TO i_{s0}
$2\Delta i_G$
ϕ
ϕ_{01}
$-V_{co_{II}}$
V_{co_I}
V_c
$V_{co_{III}}$
i_G
CHANGE OF SUPPLY CURRENT FROM i_{s0} TO $-i_{s0}$
CHANGE OF SUPPLY CURRENT FROM $-i_{s0}$ TO i_{s0}
(A)
i_s
i_{s0}
$-i_{s0}$
$\frac{V_{co_I}C}{i_0-\Delta i_G}$
$\frac{V_{co_I}C}{\sqrt{i_0^2-\Delta i_G^2}}$
i_G
$i_0-\Delta i_G$
$i_0+\Delta i_G$
t
V_c
V_{co_I}
$V_{co_{II}}$
$\frac{V_{co_I}}{V_{co_{III}}}=\frac{V_{co_I}}{V_{co_{II}}}\times\frac{V_{co_{II}}}{V_{co_{III}}}=\frac{i_0+\Delta i_G}{i_0-\Delta i_G}$
ϕ_{01}
ϕ_{02}
(B)

Fig. 6 (left). Operation at R=0 with optimum input frequency divided by 2

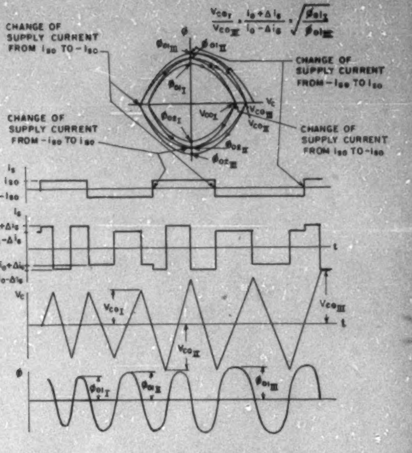

$\frac{V_{co_I}}{V_{co_{II}}}=\frac{i_0+\Delta i_s}{i_0-\Delta i_s}=\sqrt{\frac{\phi_{01_I}}{\phi_{01_{II}}}}$
CHANGE OF SUPPLY CURRENT FROM i_{s0} TO $-i_{s0}$
ϕ
$\phi_{01_{III}}$
$\phi_{01_{II}}$
ϕ_{01_I}
CHANGE OF SUPPLY CURRENT FROM $-i_{s0}$ TO i_{s0}
CHANGE OF SUPPLY CURRENT FROM $-i_{s0}$ TO i_{s0}
ϕ_{02_I}
V_{co_I}
V_c
$V_{co_{III}}$
$V_{co_{II}}$
CHANGE OF SUPPLY CURRENT FROM i_{s0} TO $-i_{s0}$
$\phi_{02_{II}}$
$\phi_{02_{III}}$
i_s
i_{s0}
$-i_{s0}$
i_s
$i_0+\Delta i_s$
$i_0-\Delta i_s$
$-i_0+\Delta i_s$
$-i_0-\Delta i_s$
t
V_c
V_{co_I}
$V_{co_{III}}$
$V_{co_{II}}$
t
ϕ
ϕ_{01_I}
$\phi_{01_{II}}$
$\phi_{01_{III}}$

Fig. 7 (right). Frequency multiplication by 2½ shown for zero losses, R=0

$$V_c=-R(i_0+\Delta i_G)\pm\sqrt{[R(i_0+\Delta i_G)]^2+\frac{2N_G}{c}(\phi_{01}-\phi)(i_0+\Delta i_G)}\quad\text{for }\phi>0$$

$$V_c=-R(i_0+\Delta i_G)\pm\sqrt{[R(i_0+\Delta i_G)]^2+\frac{2N_G}{c}(\phi-\phi_{02})(i_0+\Delta i_G)}\quad\text{for }\phi<0\qquad(9)$$

If Δi_G is exchanged by $-\Delta i_G$ in equation 9, the trajectories for $i_s=-i_{s0}$ are obtained. The two trajectories in equation 9 should be continuous at $\phi=0$. At $R=0$, they become parabolas, as in equation 4 and Fig. 4. At $R\neq0$, they combine to a kind of spiral as in Fig. 5. For R small, these approximations are made:

1. The V_{c0} decreases by $4R(i_0\pm\Delta i_G)$ per cycle of the gate circuit relaxation oscillations. This is stated in the form

$$V_{c0\,II}=V_{c0\,I}-4R(i_0\pm\Delta i_G)\qquad(10)$$

The + and − signs hold for $i_s=i_{s0}$ and $i_s=-i_{s0}$, respectively.

$V_{c0\,I}$, $\phi_{01\text{ or }2I}$, $V_{c0\,II}$, and $\phi_{01\text{ or }2II}$, are the values of V_{c0} and $\phi_{01\text{ or }2}$ in successive cycles of the gate circuit oscillation, respectively.

2. The decrease in the maximum flux per gate circuit cycle can be approximated by equations 9 and 11

$$|\phi_{01\text{ or }2\,II}|=|\phi_{01\text{ or }2\,I}|-\frac{8R^2(i_0\pm\Delta i_G)c}{N_G}\qquad(11)$$

The + and − signs hold for $i_s=i_{s0}$ and $i_s=-i_{s0}$, respectively.

3. The cycle period of the gate circuit oscillations also decreases approximately by $4Rc$ per cycle. This is shown by substituting equation 9 in 8 and integrating

$$T_{n\text{th cycle}}=T_{(n+1)\text{th cycle}}+4Rc\qquad(12)$$

Balance of Input and Output Powers

The effects of changing supply current i_s from i_{s0} to $-i_{s0}$, or vice versa, is considered next. Fig. 6 illustrates the circuit operation for four changes in supply current per cycle of the oscillations in the gate circuit.

If τ denotes the phase in the cycle of gate circuit oscillation at which the supply current changes polarity, $\tau=0$ corresponds to the instant when $V_c=0$ and $\phi=\phi_{01}$ or $\phi={}_0\phi_2$. If ϕ_c is the flux at the instant of changing the supply current i_s, the dependence of ϕ_c on τ for $R=0$ can be found by integration of equation 3 for the period of from $t=T/4$ to $t=T/4+\tau$ at $t=T/4$, $V_c=0$, and $\phi=\phi_{01\text{ or }2}$ and $\tau=0$. This results in

$$|\phi_c|=\frac{1}{2N_Gc}(i_0\pm\Delta i_G)\left[\left(\frac{T}{4}\right)^2-\tau^2\right]\quad\text{for }-\frac{T}{4}<\tau<\frac{T}{4}\qquad(13)$$

The + and − signs hold for changes of supply current from $+i_{s0}$ to $-i_{s0}$ and from $-i_{s0}$ to $+i_{s0}$, respectively. Substituting the value of T, from equation 6, using the equation at $i_s=-i_{s0}$ for a change from $-i_{s0}$ to i_{s0}, and the equation at $i_s=-i_{s0}$ for a change from i_{s0} to $-i_{s0}$, equation 13 can be written

$$|\phi_c|=\frac{1}{2N_Gc}(i_0\pm\Delta i_G)\left[\left(\frac{V_{c0}c}{i_0\pm i_G}\right)^2-\tau^2\right]\qquad(14)$$

$\tau=0$ corresponds to two points in the $V_c-\phi$ plane, i.e., either to $\phi=\phi_{01}$ or $\phi=\phi_{02}$. Therefore, τ can be defined everywhere if it is defined in the region of $-T/4<\tau<T/4$ only; ϕ and V_c must be continuous in respect to time. Hence, the two trajectories of the families Γ_1 and Γ_2 which are interchanged at the instant of supply current change should intersect at the point in the plane $(\phi=\phi_c)$ which corresponds to the instant of the change (τ). Matching the two trajectories at this point, the relationship between $V_{c0\,I}$ and $V_{c0\,II}$ is found: the voltages at $\phi=0$, in the cycles before and after the change in supply current, respectively. From equations 4 and 14, this relationship is found to be

$$V_{c0\,II}^2=V_{c0\,I}^2\pm\frac{4N_G}{c}\Delta i_G|\phi_c|\qquad(15)$$

where the + and the − signs hold for changes of the supply current from $-i_{s0}$ to $+i_{s0}$ and from $+i_{s0}$ to $-i_{s0}$, respectively.

The exchange of energy with the change of supply current is determined from equation 15 or from considerations of the stored electric or magnetic energy. The change of the supply current, as viewed from the gate circuit, is regarded as a change of the characteristic of the nonlinear inductance, as shown in Fig. 3.

As ϕ cannot change instantaneously, the gate circuit current has to change instantaneously by $2\Delta i_G$, along a horizontal line in the ϕ versus i_G plot. Fig. 3(D), the energy inserted, is found from the change of stored magnetic energy in the gate circuit

$$e=\left[\int_0^{\phi_c} i_G d\phi\right]_{\text{After change in supply current}} - \left[\int_0^{\phi_c} i_G d\phi\right]_{\text{Before change in supply current}} \quad (16)$$

Thus the energy exchanged can be represented by one of the shaded areas in Fig. 3(D). Hence

$$e=\pm 2N_G\Delta i_G|\phi_c| \quad (17)$$

where the + sign holds for a change in supply current from $-i_{s0}$ to $+i_{s0}$ and the − sign holds for a change in supply current from $+i_{s0}$ to $-i_{s0}$. Energy is added at a change from $-i_{s0}$ to $+i_{s0}$ and is subtracted at a change in supply current from $+i_{s0}$ to $-i_{s0}$. In order to have a maximum power input into the gate circuit, the change from $-i_{s0}$ to $+i_{s0}$ should occur with ϕ_c as close as possible to ϕ_{01} or ϕ_{02}; and from i_{s0} to $-i_{s0}$ should occur with ϕ_c as close as possible to zero. In the ideal case, $\phi_c=\phi_{01}$ or $_2$ for a change from $-i_{s0}$ to $+i_{s0}$ and $\phi_c=0$ for a change in supply current from $+i_{s0}$ to $-i_{s0}$. There is then one energy impulse added per cycle of the supply current.

To obtain steady state operation, a balance of input energy and losses must be maintained. The power input can be written

$$P_{\text{input}}=\sum_{f_0\text{ cycles}} 2N_G\Delta i_G\times \left[\phi_{c(\text{change from } -i_{s0} \text{ to } +i_{s0})}-\phi_{c(\text{change from } +i_{s0} \text{ to } i_{s0})}\right] \quad (18)$$

The energy losses go into the load and into the circuit resistive and hysteresis loop losses. The losses have been considered previously as consisting of a resistance in the gate circuit only (i.e., i_G^2R). Some remarks about the simplifying assumption made at the beginning of this paper can now be made through an examination of Fig. 3(D) and equation 16.

In the case of a square-wave current supply, the change from the characteristic for $i_s=i_{s0}$ to that of $i_s=-i_{s0}$ is very rapid. Therefore, the horizontal line of $\phi=\phi_c$ represents this change. The shaded area represents then the energy transferred. If the change from $i_s=i_{s0}$ to $i_s=-i_{s0}$ is gradual, as in the case of i_s having a sinusoidal waveshape, the line referring to this in the ϕ versus i_G plane is shown in Fig. 3(D) inside the shaded area. The area enclosed between this line and the i_G axis represents the energy transmitted, which is smaller than the energy for a square wave.

The error involved in assuming zero flux change in the flux-saturated region can now be assessed. A more realistic characteristic than the one in Fig. 3(A) is now assumed, the flux in the saturated region varying linearly with the current. Following a construction similar to the one leading to Fig. 3(D), the area in the ϕ versus i_G plot, representing energy transmitted, is reduced by a ratio equal to the ratio of the flux change in the flux-saturated region to the flux change ϕ_c in the nonsaturated region.

The waveshape of the gate circuit current is repeated after an even number of supply current cycles. The balance of energy should be calculated for the duration of the minimum even number of cycles of the supply current in which the gate circuit wave shape is not repeated.

To illustrate further this form of operation, two modes have been described for $R=0$, in Figs. 6 and 7. In Fig. 6, the frequency of the gate circuit oscillation is half that of the supply current. In Fig. 7, frequency of gate circuit oscillations is $2^1/_2$ times greater than the supply current frequency. The wave shape of the gate circuit oscillation is repeated every two cycles of supply current. In both Figs. 6 and 7, operation is described for no losses in the gate circuit and optimum conditions of power input

$$\phi_{c(\text{change from } -i_{s0} \text{ to } i_{s0})}=\phi_{01 \text{ or } 02}$$
$$\phi_{c(\text{change from } +i_{s0} \text{ to } -i_{s0})}=0$$

Under these conditions, the energy stored in the gate circuit increases each cycle of the supply current as is evident from the consistent increase of V_{c0} and ϕ_{01} or $_2$. Correspondingly, the supply current period would have to increase, too. Since the latter is impractical, balance of energy is one condition for stability. To realize steady state operation, input into the gate circuit should be balanced by the losses.

Stability

The foregoing discussion is sufficient for deriving certain conclusions outlined herewith.

"Off" mode: If $\phi=0$ and $\phi_c=0$, there is no input energy, equation 17, and only slight oscillations in the gate circuit.

"On" mode: In order to switch the circuit to an "On" mode, these three requirements should be satisfied:

1. An impulse of energy is required that will bring the gate circuit to a point in the $\phi-V_c$ plane on or outside a trajectory in which $V_{c0\,I}$, $V_{c0\,II}$ and ϕ_c, from equations 15 and 17, will allow sufficient energy input from the supply circuit to balance the losses. This requirement will allow the point in the $\phi-V_c$ plane to be in any one of the four quarters of the plane.
2. At the time the switching impulse of energy is applied, the phase τ in respect to the supply current oscillation should be such that energy will be added in each cycle of supply circuit current. The condition for positive energy input from equations 16, 17, and 18 are that approximately for

$$\frac{\Delta i_G}{i_0}\ll 1$$

from equation 7

$$T(i_s=+i_{s0})\approx T(i_s=-i_{s0})- \qquad \frac{T}{8}<\tau<\frac{T}{8} \quad (19)$$

In Fig. 8, the two lines, representing loci of the point where $\tau=T/8$ and $\tau=-T/8$, are drawn in the $\phi-V_c$ plane. The quarters of the plane, where there is positive energy input, are so marked.

3. The third requirement is stability.

While oscillations can be maintained under balanced conditions of input and output power, the question now is whether this balance is stable. Assuming that a small disturbance is introduced in timing the supply current and gate circuit oscillations, the balance of power would then be disturbed. Stability would require that, subsequent to the introduction of the disturbances, the operation would return automatically to the initial timing and amplitudes that correspond to the power balance conditions.

A steady state operation with the output and input powers balanced—gate circuit frequency being one-half that of supply circuit frequency—is described in Fig. 6 where the trajectory in the $\phi-V_c$ plane circles the origin once per two cycles of the supply current.

Changes of supply current occur at $\phi=\phi_c$ and at phase τ. A disturbance is introduced in the form of a change of supply circuit current at $\phi_c=\phi_c+\Delta\phi$ and $\tau'=\tau+\Delta\tau$, instead of at ϕ_c and τ which are required by the balance of energy; $|\Delta\phi|$ and $|\Delta\tau|$ will be examined in subsequent cycles. If $\Delta\phi$ and $\Delta\tau$ tend to zero or remain small, the operation will be stable; otherwise, it will be unstable.

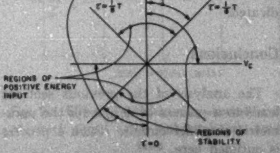

REGIONS OF POSITIVE ENERGY AND STABILITY
φ
τ=0
DIRECTION OF TRAJECTORIES IN PLANE
τ=⅛T
τ=⅛T
V_c
REGIONS OF POSITIVE ENERGY INPUT
REGIONS OF STABILITY
τ=0

Fig. 8. Indication of regions in the $\phi-V_c$ plane where conditions of positive energy input and stability are satisfied

As $\Delta\tau$ and $\Delta\phi$ are related, equation 14, it is sufficient to analyze the behavior of $\Delta\tau$. The change in τ affects the oscillation periods in the gate circuit.

The circuit operation in the "on" mode will be stable if $|\Delta\tau|$ will consistently decrease each cycle of the supply frequency.[2] The condition of stability can then be determined. In case that $\Delta\tau$ is positive, the period of the gate circuit oscillation, T, should decrease so that $\Delta\tau$ in the successive cycle will decrease as well. If $\Delta\tau$ is negative, T should increase so that $\Delta\tau$ can decrease. This condition is stated as

$$\frac{dT}{d\tau}<0 \tag{20}$$

On the other hand, the change in T should not be so large as to cause $\Delta\tau$ to increase in the opposite polarity. If the phase error is $\Delta\tau$ for the nth cycle, it should not decrease by more than $2\Delta\tau$ to have the phase error in the $n+1$ cycle more negative than $-\Delta\tau$. This condition is stated as

$$\frac{dT}{d\tau}>-2 \tag{21}$$

From equations 6 and 13 it is possible to see that the condition in equation 20 is satisfied for the first and third quarters, as shown in Fig. 8, of the $\phi-V_c$ plane; namely, for

$$0<\tau<\frac{T}{4} \tag{22}$$

The condition in equation 21 may fall in the range where the power input is negative and would not impose additional limitation on the range of operation.

Fig. 8 indicates parts of the $\phi-V_c$ plane in which both stability and positive energy input are possible, corresponding to

$$0<\tau<\frac{T}{8} \tag{23}$$

Switching of the circuit into one of two "on" modes—each at 180-degree phase difference from the other—is possible by applying an energy impulse of sufficient magnitude and timing as indicated.

Conclusions

The analysis of the previous sections leads to a process of design for the parametric amplifier circuit. Such a process is outlined here.

Data such as enumerated are usually available to the designer prior to initiating the design:

1. Average load current—i_0.

2. Resistance representing load and internal losses—R.

3. Frequency and period of supply current, f_0 and T_0—twice the output frequency of the parametric amplifier oscillations.

4. In order to proceed with the design, an assumption of the ratio, $\Delta i_G/i_0$ is necessary. The analysis in the previous sections describes a mode of operation where $\Delta i_G/i_0 \leq 1$.

Practically, because of tolerances, imperfect magnetic characteristic, etc., the ratio $\Delta i_G/i_0$ should be less than one-half. In the discussion on stability, the periods for the four quarters of the oscillation cycles were assumed to be the same. This assumption can be justified for $\Delta i_G/i_0 \approx 1/4$. If this ratio is not maintained $T(i_s=i_{s0})/T(i_s=-i_{s0}) \ll 1$; see equation 7. This, in turn, narrows the angle corresponding to positive power input in the $V-\phi$ plane; Fig. 8.

The period of the pump frequency T_0 is

$$T_0=1/4[T(i_s=i_{s0})+T(i_s=-i_{s0})] \tag{24}$$

Because of remitting from oscillations at one-half the supply frequency symmetry, assume that $V_{c0}(i_s=-i_{s0})=V_{c0}(i_s=+i_{s0})$. From equations 24 and 7, $T(i_s=i_{s0})$ and $T(i_s=-i_{s0})$ can be calculated. For $\Delta i_G/i_0 \ll 1$ from equation 7: $T(i_s=i_{s0}) \approx T(i_s=-i_{s0})=2T_0$.

Next τ is selected. At $\tau=0$, the energy input is maximum but the circuit borders on the unstable region; see Fig. 8. At $\tau=T/8=T_0/4$, the power input is zero, and τ should be chosen in the range between these two extreme values. The middle of this region, $\tau=T/16=T_0/8$, gives a compromise between stability and power input.

N_G, ϕ_{01}, or ϕ_{02}, and c can be determined from two conditions. One is the timing condition, similar to the one expressed for $R=0$ in equation 6. This condition should provide the value of the product $V_{c0}c$. The second condition is that of balance input and output powers as expressed in equation 18. Substituting equation 14 in 18 to eliminate ϕ_c, the later condition should provide the value of $V_{c0}c$. An iterative method of solution is adopted where the zero approximation corresponds to $R=0$. Improved approximations are arrived at, making use of equations 9, 10, and 11 where $R\neq 0$.

Circuits are arranged in tandem, driving one another.[1] In this manner of operation, bias and alternating supply currents are supplied initially to the driving circuits but not to the driven circuit. A number of driving circuits over a number of cycles build up the energy stored in the driven circuit. The driven and driving circuits oscillate then at the same frequency, with a phase difference equivalent to the delay caused by the connecting wires. If the distance between the two circuits is l feet, the phase difference is approximately l millimicroseconds. When the energy in the driven circuit is sufficiently high (conditions of turning "on" as defined previously), the supply current is applied to the driven circuit. When the driving circuits are turned "off" by removal of supply current, the driven circuit continues to oscillate if the phase satisfies the stability condition. If the phase of the driving circuit is, as assumed above, $=T_0/8$, in order to obtain stable oscillations in the driven circuit

$$\frac{T_0}{8}\text{ (in millimicroseconds)} \leq l \text{ feet}$$

Nomenclature

c = capacitance
e = energy input per half-cycle of supply frequency
f_0 = frequency
i_G = gate circuit current
i_{s0} = amplitude of square-wave supply circuit current
Δi_G = change in gate circuit current that corresponds to i_{s0}
l = distance in feet between driven and driving circuits
N_B, N_G, N_S = number of turns in bias, gate, and supply circuits windings, respectively
R = gate circuit resistance
P = power
T = periods of gate circuit oscillation
T_0 = period of supply frequency
t = time
V_c = voltage across capacitor
V_{c0} = voltage across capacitor at $\phi=0$
$V_{c0\ I}$, $V_{c0\ II}$, etc. = values of V_{c0} in successive cycles before and after a change in supply current
ϕ = flux
ϕ_c = flux at change of supply current
$\Delta\phi_c$ = deviation from ϕ_c caused by Δt
ϕ_s = saturation flux
ϕ_{01}, ϕ_{02} = maximum positive and negative fluxes, respectively, at $V_1=0$
$\phi_{0(1\text{ or }2)I}$, $\phi_{01(1\text{ or }2)II}$, etc. = $\phi_{01\text{ or }2}$ in successive cycles of gate circuit current; also, before and after a change in supply circuit current
ϕ_1, ϕ_2 = fluxes in cases 1 and 2
τ = time, indicating phase
$\Delta\tau$ = deviation from τ in stability examination

References

1. THE PARAMETRON, A DIGITAL COMPUTING ELEMENT WHICH UTILIZES PARAMETRIC OSCILLATION, Eiichi Gota. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 47, no. 8, Aug. 1959, pp. 1304–16.

2. FREQUENCY MULTIPLICATION WITH CURRENT SATURABLE INDUCTANCE, N. S. Prywes. *Ph.D. Thesis*, Harvard University, Cambridge, Mass., 1954.

Optical Read-out of Digital Magnetic Recording

J. J. MIYATA
NONMEMBER

THE theoretical aspects of saturation magnetic recording have been under investigation at the Electronics Division of The National Cash Register Company for several years. One conclusion of this study pointed out a significant difference between tape and drum (or disk) systems in regard to their storage density limitations.

Limitation of Present Techniques

Neglecting mechanical limitations of the handler, the major limitation in linear storage density on tape, where the head is in contact with the coating, is attributable to the recording medium itself. See Fig. 1. On the other hand, the chief limitation for drum or disk memories, in which the head does not contact the coating, is caused by the separation of the playback head when the separation is approximately 0.0002 inch or greater. This limitation becomes more pronounced as the resolution characteristics of the coating are improved.

Figs. 2 and 3 show the effect of this separation on the playback signal from a high-resolution oxide coating and an experimental high-density magnetic coating. With a conventional playback head, the output signal is a series of pulses corresponding to the flux reversals in the coating. The pulse width determines the maximum storage density possible,[1] and pulse width is defined here as the distance along the coating over which the signal amplitude exceeds 20% of the peak value.

The curves show that the playback head separation is primarily responsible for loss in resolution and amplitude. Deterioration in signal is inevitable as the separation increases, because the signal from a conventional playback head is sensitive to the change rate of the flux at some distance from the coating. Playback head limitation is inherent in any type of head which requires external flux linkages from the coating to generate a signal. Direct head-coating contact is undesirable in a drum or disk system because head and coating wear may lead to unreliable performance.

A playback head-to-coating separation of only 0.00005 inch with an oxide coating will cause a 15% increase in pulse width. A playback head separation of 0.001 inch will increase the pulse width 400%. This resolution loss becomes more pronounced when a special high-density coating is used.

Only a radically different approach to the playback head design can lead to a "breakthrough" which will eliminate this basic limitation. One approach is to use a light beam to "read" the surface magnetization of the coating without making mechanical contact with it, thus eliminating head and coating wear without resolution loss. Fig. 4 compares the density to be expected from such a system and that of a conventional system using a regular read-back head with an oxide coating. The density is normalized to that of an in-contact system using a high-resolution oxide coating. The optical read-out head makes use of the magneto-optic effect.

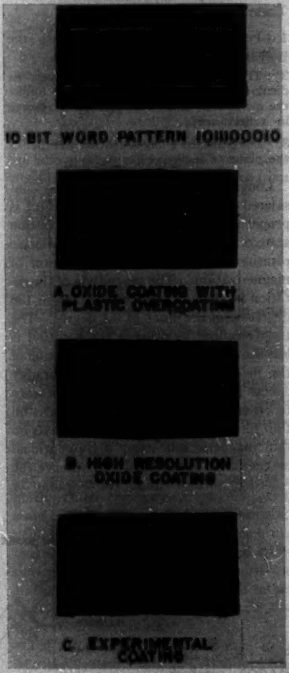

10 BIT WORD PATTERN 1011100010
A. OXIDE COATING WITH PLASTIC OVERCOATING
B. HIGH RESOLUTION OXIDE COATING
C. EXPERIMENTAL COATING

Fig. 1. Dependence of playback signal on magnetic coating (1,000 bits per inch NRZ)

The Magneto-Optic Effects

Magneto-optic effects are created by interaction between light and matter when matter is magnetized. The phenomenon may manifest itself in the light as a change of wavelength accompanied by polarization (Zeeman effect); a magnetic double refraction (Voigt and Cotton-

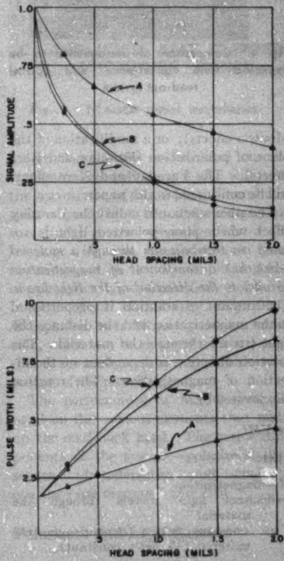


Fig. 2. Effect of head-coating separation on playback signal using high-resolution oxide tape

A—Read head in contact
B—Record head in contact
C—Both heads out of contact

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 13, 1960; made available for printing September 29, 1960.

J. J. MIYATA is with The National Cash Register Company, Hawthorne, Calif.

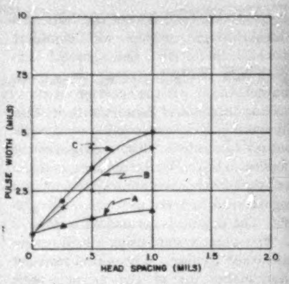


Fig. 3. Effect of head-coating separation on playback signal using experimental high-density coating

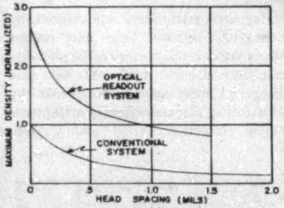


Fig. 4. Comparison of performance to be expected from conventional and optical read-out system

Mouton effects); or a modification of the plane of polarization (Faraday and Kerr effects). The Faraday and Kerr effects will be considered in this paper.

The phenomenon is called the Faraday effect when plane-polarized light is rotated on *transmission through a material which has a component of magnetization parallel to the direction of the light beam.* The amount of rotation is proportional to the magnetization and the distance the light travels through the material. The rotation direction is dependent on the direction of magnetization. The rotation is expressed by

$$\theta = VMl$$

where

M = magnetization
l = distance light travels through the material
V = a constant which depends on the material (Verdet's constant)

When the plane of polarization is altered upon *reflection from a magnetized surface,* the phenomenon is called the Kerr effect. The reflected light is, in general, elliptically polarized, but the degree of ellipticity is small so that, to a first approximation, the effect will be regarded as a rotation of the polarization plane. In the Kerr effect, there are three possible orientations of the magnetic field with respect to the plane of incidence. These effects are called:

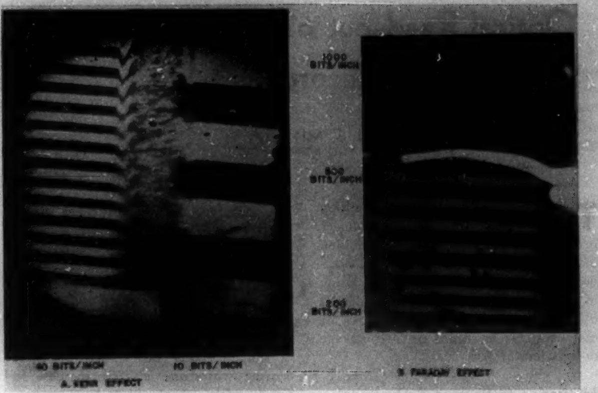


Fig. 5. Magnetically recorded patterns using Kerr and Faraday effects

1. Polar (magnetization normal to the reflecting surface).
2. Transverse (magnetization perpendicular to the plane of incidence; unlike the other two, this effect produces a change in reflectance).
3. Longitudinal (magnetization parallel to the plane of incidence).

Longitudinal effects only will be considered in this paper. The Kerr effect is proportional to the net magnetization state of the material with the direction of rotation dependent on the direction of magnetization. The magnitude of the rotation in changing the magnetization 180 degrees is of the order of 0.1 degree. The time required to observe the rotation is less than 10^{-8} seconds after application of the field. Both the Kerr and Faraday effects have been used in the laboratory to observe magnetic domain patterns on thin films.[2-4] One can record signals on thin magnetic film using a conventional magnetic recording head and observe the recorded pattern, using the Faraday or Kerr effects. Fig. 5 shows photographed signals of these effects.

Magneto-Optic Playback Head

The basic system using the Kerr effect to read out magnetically recorded signals is shown in Fig. 6. The laboratory model to prove the feasibility of the magneto-optic playback head (MAGOP) is shown in Fig. 7. This paper discusses the Kerr effect in MAGOP. The Faraday effect analysis would be similar.

The light from the source is plane-polarized and concentrated onto the rotating magnetic disk. The reflected light is collected by the lens and passed through the analyzer onto a photodetector, which converts the light signal into an electric one. By proper alignment of polarizer and analyzer, the positive magnetization state ("one" bit) can be made to correspond to the higher transmission state of the analyzer and the negative magnetization ("zero" bit) to the lower state. The plane of polarization of the incident light must be parallel or perpendicular to the plane of incidence in order to eliminate the ellipticity caused by metallic reflection.

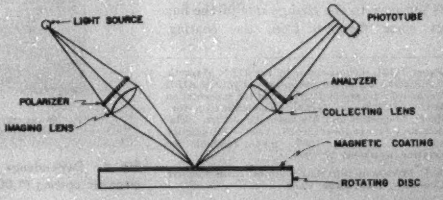


Fig. 6. Basic MAGOP playback head using Kerr effect

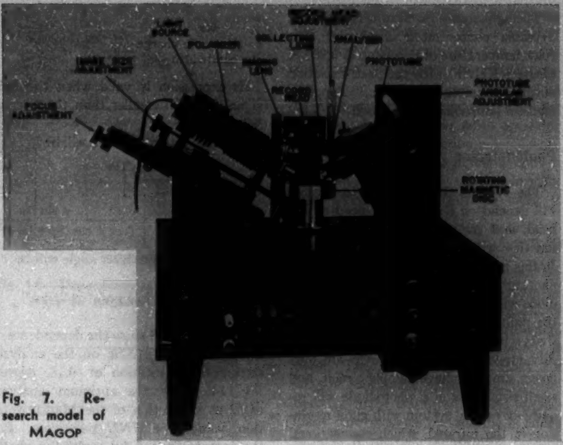


Fig. 7. Research model of MAGOP

When plane-polarized light is incident on the analyzer, the light transmitted will be proportional to $\cos^2\phi$ where ϕ is the angle between the polarizer and analyzer axes. Thus:

$$L_+ = L_0t[\cos^2(\phi+\psi_+)+c]$$

$$L_- = L_0t[\cos^2(\phi+\psi_-)+c]$$

where

L_+ = light transmitted through the analyzer when the surface is positively magnetized
L_- = light transmitted when the surface is negatively magnetized
L_0 = intensity of the light source
t = transmission of the optical system with the axes of the polarizer and analyzer aligned
ψ_+ = magneto-optic rotation caused by positive magnetization
ψ_- = magneto-optic rotation caused by negative magnetization
c = term due to imperfect polarization

Assume that

$$\psi_+ = -\psi^- = \psi/2$$

where

ψ = total rotation

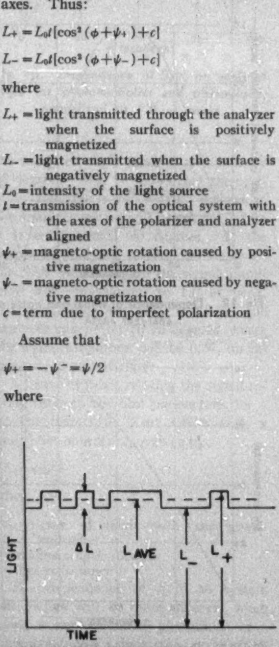


Fig. 8. Light transmitted by analyzer

The light signal will be

$$\Delta L = L_+ - L_- = L_0t\,|\sin\psi\,\sin 2\phi|$$

In general the magneto-optic rotation will be small so

$$\Delta L \cong L_0t\psi\,|\sin 2\phi|$$

The light signal (ΔL) is proportional to the above rotation, ψ, and is a maximum when the analyzer axis is 45 degrees from that of the polarizer.

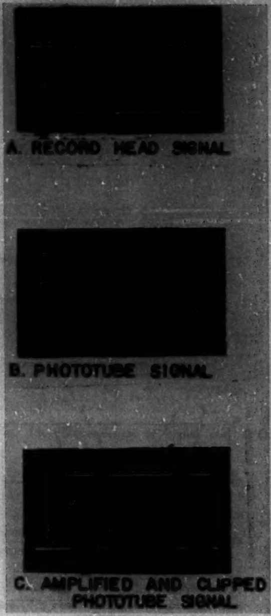
The average light level is given by

$$L_{\text{avg}} = \frac{L_0t}{2}(\cos\psi\,\cos\,\phi + 1 + c)$$

Thus, the light transmitted by the analyzer will be composed of a high average light level on which is superimposed the useful MAGOP signal; see Fig. 8.

The useful MAGOP signal, ΔL, will have the same waveshape as the record current (saturation recording) with the amplitude of the signal independent of the rotation speed. This head may be classified as a flux sensitive head, whose output is proportional to the surface magnetization state of the coating. It differs from conventional flux sensitive heads, which use balanced magnetic circuits or Hall effect devices, in that the output is proportional to the surface magnetization state rather than to the stray flux field at some distance from the coating.

Fig. 9 shows the waveshape in a MAGOP system. Fig. 9(A) shows the current waveshape in the record head. The magnetization of the coating will follow that of the record head current (saturation recording) so it also will have the waveshape of Fig. 9(A). The phototube output Fig. 9(B) will have the same waveshape as the magnetization of the coating with the photodetector noise superimposed on it. By amplifying and clipping this signal, a relatively noise-free output signal is obtained, Fig. 9(C). Except for an arbitrary polarity inversion, which can be changed with the position of the analyzer, this signal is identical to that of the record signal.


A. RECORD HEAD SIGNAL

B. PHOTOTUBE SIGNAL

C. AMPLIFIED AND CLIPPED PHOTOTUBE SIGNAL

Fig. 9. MAGOP signal waveshapes

The performance of MAGOP will depend on the various elements that make up the read-back head. The major components of the system—polarizers, light source, photodetector, and magnetic coating—are discussed herewith.

Polarizers and Light Sources

The polarizers for the MAGOP are made in sheet form by Polaroid Corporation. This type offers low cost and minimum loss in image quality. The choice of specific polarizers depends on the photodetector. For visible light operation, the *HN22*, which transmits 0.0005% with

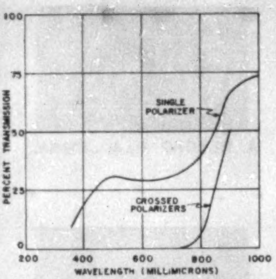


Fig. 10. Spectral characteristics of HN22 polarizer

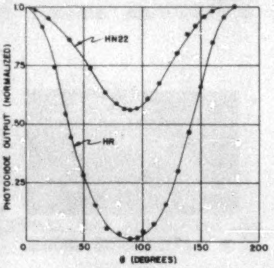


Fig. 11. Comparison of visible and infrared polarizer efficiencies in near infrared

axes crossed, may be used,[5] and this type is effective in polarizing light out to approximately 750 millimicrons; see Fig. 10. The near infrared region requires type *HR* infrared polarizers, which cover the spectral band from approximately 700 millimicrons to almost 3 microns.

Since, in general, the sensitivity of the multiplier phototubes do not extend beyond 750 millimicrons, *HN22* polarizers are adequate. The semiconductor photodiodes and photoconductors have sensitivities in the visible as well as the infrared. With these detectors, *HR* polarizers must be used, the *HN22* polarizers being ineffective in the infrared. Fig. 11 shows a comparison of the efficiencies of the *HN22* and *HR* polarizers when using a germanium photodiode with a Sylvania Electric Products, Inc., 25-watt concentrated arc lamp. The extinction values are 56% and 0.25% for crossed *HN22* and *HR* polarizers. Better polarization can be obtained with the *HN22* by using a filter to eliminate the radiation beyond 750 millimicrons; however, most of the effective radiation will then be filtered out.

The light source for MAGOP is a 25-watt Sylvania concentrated arc lamp. The color temperature of this lamp is 3,200 K (degrees Kelvin) with the peak radiation at approximately 1 micron. Tungsten filament lamps may also be used.

Photodetector

The type of detector used in MAGOP will depend on restrictions placed on the head, and the performance and characteristics will depend on the detector. In this paper, operation with a multiplier phototube and a semiconductor photodiode will be considered. Advantages of the multiplier phototube are high sensitivity, high output, and fast response. Advantages of the photodiode are low-power supply voltage, simple circuit, and small size. Use of a photodiode will require an optical system with good resolution in the infrared, which conventional glass optics generally do not possess.

Two types of multiplier phototubes and seven types of photodiodes were tested for MAGOP and the best photomultiplier and photodiode were used in the MAGOP studies. Photodiodes requiring cooling were not considered because of the added complexity of the cooling device.

Magneto-optic rotation is generally small (≈ 0.10 degree) so the photodetector and film noise may be limiting factors in obtaining high-density recording. The photodetector noise characteristics are quite different for the multiplier phototube and the photodiode. The film noise may result from either a polarization change or reflectance change.

Photodetector Noise

The dependence of signal-to-noise ratio (S/N) in a system in which the detector noise is predominant will depend on the type of detector used. In the case of the playback head with a multiplier phototube as the detector, the signal amplitude will be given by

$$S=\eta A^m R\Delta L=\eta A^m R L_0 l\psi\,|\sin 2\phi|$$

where

S = signal voltage
η = cathode sensitivity
A = amplification per dynode
m = number of dynodes
R = load resistance

The noise voltage from the multiplier phototube, where the predominant noise results from photoemission shot noise, is given by

$$N=A^m R\left[\frac{eB\eta L_0 lA}{A-1}(\cos\psi\cos 2\phi+1+c)\right]^{1/2}$$

where

e = electronic charge
B = bandwidth

This expression is valid when the light signal is much smaller than the average light level.

The signal-to-noise ratio will be

$$(S/N)=\left[\frac{\eta L_0 l(A-1)}{eBA}\right]^{1/2}\times\frac{\psi\sin 2\phi}{(\cos\psi\cos 2\phi+1+c)^{1/2}}$$

The optimum analyzer angle will be

$$\phi_m=\frac{1}{2}\cos^{-1}\left[-(c+1)+\sqrt{c(c+2)+\psi^2}\right]$$

Figs. 12 and 13 show the dependence of signal, noise, and S/N on the analyzer setting for a rotation of 0.10 degree. Fig. 14 shows the optimum analyzer angle as a function of the MAGOP rotation, using type *HN22* polarizers. For the system limited by multiplier-phototube noise, the optimum angle will be very near the extinction angle. For example, a film giving a rotation of 0.1 degree will have the maximum S/N with

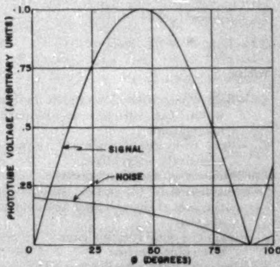


Fig. 12. Dependence of signal and noise on analyzer angle

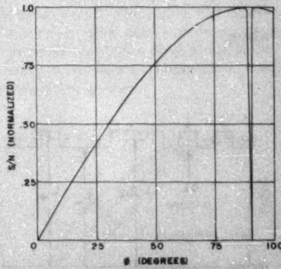


Fig. 13. Dependence of S/N on analyzer angle

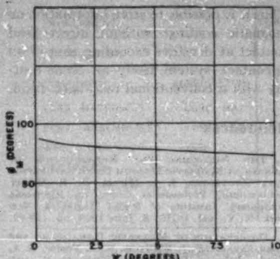


Fig. 14. Dependence of optimum analyzer angle on magneto-optic rotation

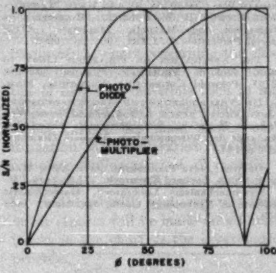


Fig. 15. Dependence of S/N on analyzer angle for photomultiplier and photodiode, normalized to 1.0 at peaks

the analyzer angle set at approximately 88 degrees. Fig. 14 shows that the optimum analyzer is not sharply defined so the analyzer setting will be critical only if it is set too near 90 degrees. At the optimum analyzer setting (88 degrees), the *S/N* will be 10 and the signal amplitude 0.06—normalized to 1.0 at peak. By sacrificing just 5% in *S/N*, the analyzer can be set to 73 degrees, where the signal amplitude will be 0.57, an increase of almost 10 times.

A figure of merit relating the signal-to-noise ratio to the film parameters, for a system using the Kerr effect with a multiplier phototube, is given by

$$M_m = \psi\sqrt{r} \qquad (1)$$

where

M_m = figure of merit with the photomultiplier
ψ = magneto-optic rotation
r = film reflectance

The dependence of *S/N* on system parameters will be quite different when using a semiconductor photodiode or a photoconductor, since the noise from these detectors is relatively independent of the incident light level. Fig. 15 shows the *S/N* as a function of the analyzer angle for a photomultiplier and a semiconductor photodiode. In actual practice the optimum analyzer angle cannot be used with the photodiode since the film noise becomes great when operating at angles far from extinction.

Since the photodiode detector noise is relatively independent of the incident light level, the film figure of merit is given by

$$M_p = \psi r \qquad (2)$$

Equations 1 and 2 demonstrate that the film figure of merit is proportional to the square root of the reflectance if a photomultiplier is used, but it is proportional to the reflectance itself if a photodiode is used. Thus, in evaluating films for MAGOP, one must consider the type of detector to be employed. This is especially true when an overcoating is added in order to increase the rotation, because this coating also acts to decrease reflectance.[6,7]

Film Noise

Film noise may be due to inhomogeneity of the film surface or to dust collecting thereon. Variation of reflectance or a change in polarization, or both, may result. If the reflected light is assumed to be plane polarized, the film noise voltage will be proportional to

$$\Delta = A\cos^2\theta - B\cos^2(\theta + E) \qquad (3)$$

where

Δ = difference between the light when it is reflected off normal and "noisy" areas
A and B = the reflectances of these areas
θ = angle between the plane of polarization and the analyzer axis
E = the change in plane of polarization at the "noisy" area

If there is no polarization change, $E = 0$ and equation 3 will simplify to

$$\Delta = (A - B) \quad \cos^2\theta$$

Thus, the noise voltage will be proportional to the difference in reflectances, and will vary as $\cos^2$ of the angle between the plane of polarization of the reflected light and the analyzer axis.

If there is no reflectance change, and only a change in polarization, $A = B$, and equation 3 will reduce to

$$\Delta \cong \frac{AE}{2}\sin 2(\theta + E/2)$$

A polarization change gives a noise voltage proportional to this change and follows a sine curve as the analyzer is rotated. The useful MAGOP signal also follows a sine curve so that noise voltage, caused by a polarization change, and the signal will vary in the same manner.

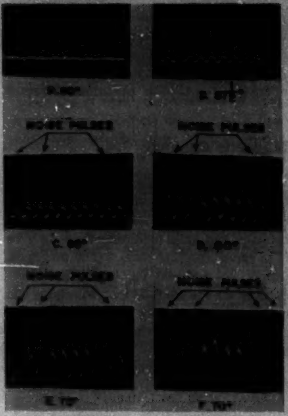


Fig. 16. Dependence of signal and film noise voltages on analyzer angle

Fig. 16 shows the effect of a small dust spot on the film, a photodiode being the detector. At angles very near extinction, the voltage caused by the dust will be much smaller than that caused by the signal. As the analyzer angle is changed, the noise voltage varies as $\cos^2\theta$ but the signal varies as $\sin 2\theta$; thus, at angles far from extinction, the cosine term predominates. Fig. 17 shows the noise pulse as a function of the analyzer angle. This curve follows the $\cos^2$ function closely, showing that the noise voltage is a result of the change in reflectance rather than in polarization. Experiments in the laboratory have shown that film noise stems primarily from reflectance rather than polarization changes.

When a photomultiplier is the detector, *S/N* is a maximum if the analyzer is set near extinction. For a semiconductor

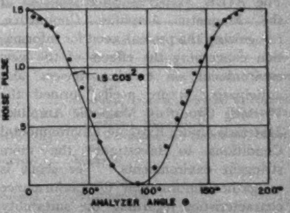


Fig. 17. Noise voltage from dust spot on film as function of analyzer angle

photodiode, the optimum occurs with the angle set at 45 degrees. Hence, film noise will be a proportionately greater problem with the photodiodes the nearer operation is to the optimum detector S/N.

Magnetic Coating

The magnetic coating for MAGOP must be very smooth and uniform to minimize noise, and produce large magneto-optic rotation and have high reflectance to give high signal-to-noise ratio. The magnetic properties required for high-density recording are square **B**-**H** loop to reduce the record head trailing effect and high coercivity to reduce the self-demagnetization effect.[1]

Various types of coatings, investigated for MAGOP, were evaporated on polished pyrex disks to give uniform surfaces. Films of iron, nickel, cobalt, and iron-nickel and iron-cobalt alloys were tested. With the angle of incidence set at 45 degrees, the largest Kerr rotation was obtained from a film of iron-cobalt alloy *(70-30)*.

The MAGOP rotation can be increased from 0.1 to 0.8 degree (double angle) by adding a dielectric zinc sulfide overcoating of the proper thickness. However, the figure of merit (M_m) of the film for a multiplier phototube increased by only 2.7 times because of the decreased reflectance.[6,7]

Conclusions

For a conventional magnetic recording system, where both record and read-back heads are separated from the coating, loss in signal resolution is caused primarily by the latter separation. Using a light beam to "read" the information will eliminate this loss since it responds to the actual surface magnetization state of the coating. The output signal will have the same waveshape as the record current and its amplitude will be independent of the speed. Since the signal depends on the coating's surface magnetization, a very thin film can be used to obtain high storage density.

The MAGOP principle of optically reading back magnetically recorded information overcomes the basic limitation inherent in any out-of-contact read-back method, which responds to a magnetic field, external to the storage surface. Thus it is possible to store information on magnetic coatings without direct head contact at densities exceeding that of an in-contact system, using an oxide coating with a conventional read-back head.

References

1. THE RECORDING AND REPRODUCTION OF SIGNALS ON MAGNETIC MEDIUM USING SATURATION TYPE RECORDING, **J. J. Miyata, R. R. Hartel.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., vol. PGEC-8, June 1959, pp. 159-69.

2. OBSERVATION OF MAGNETIC DOMAINS BY THE KERR EFFECT, **H. J. Williams, F. G. Foster, E. A. Wood.** *Physical Review*, New York, N. Y., vol. 82, 1951, p. 119.

3. MAGNETIC DOMAINS ON SILICON IRON BY THE LONGITUDINAL KERR EFFECT, **C. A. Fowler, E. M. Fryer.** *Ibid.*, vol. 86, 1952, p. 426.

4. OBSERVATION OF MAGNETIC DOMAINS AND DOMAIN WALL IN BiMn BY THE KERR MAGNETO-OPTIC EFFECT, **B. W. Roberts.** "Conference on Magnetism and Magnetic Materials," *AIEE Special Publication T-78*, Oct. 1955, pp. 192–97.

5. THE APPLICATION OF POLARIZED LIGHT TO BASIC DESIGN PROBLEMS. *Polaroid Bulletin F2165*, Polaroid Corporation, Cambridge, Mass.

6. DIE VERGROBERUNG DER MAGNETO-OPTISCHEN KERRDREHUNG DURCH AUFDAMPFSCHICHTEN, **Von Waltraut Heinrich.** *Sitzungberichten*, Bayerischer Akadamie der Wissenschaften, Münich, Germany, 1956, p. 133.

7. OPTICAL DATA-PROCESSING WITH THIN MAGNETIC FILMS, **Leonard Kleinrock.** *7848–7849-TM-4*, Servomechanisms Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., June 1959.

Report on Behavior of Magnetic Materials Under Extreme Environmental Conditions

AIEE COMMITTEE REPORT

WITH THE ADVENT of space vehicles, missiles, and outstanding developments in the aircraft and marine transportation industries came new and more stringent environmental conditions for electronic devices and equipments. The AIEE Materials Subcommittee of the Magnetic Amplifier Committee, recognizing the present need for information concerning the effects of the new environment on existing devices and anticipating future needs, formed the Working Group on Magnetic Amplifier Materials Under Extreme Environmental Conditions to investigate the more stringent environments. The scope is to gather information on performance characteristics regarding the suitability of principal components used in magnetic amplifiers such as cores, wires, rectifiers, insulation, and potting as affected by extreme environmental influences; and to publish information describing the state of the art and to point out where more work is needed.

In order to gain a better understanding of the problems associated with each environmental condition and to establish grading levels, a chart was constructed listing environmental conditions and magnetic amplifier components. The original chart was all-embracing but, after careful study, it was modified to include only those present-day areas considered to be most significant. The tentative chart adopted by the group as a guide, Fig. 1, lists environments of temperature, nuclear radiation, mechanical stress, and atmosphere. The magnetic amplifier components are core materials, core assemblies, wire conductors, rectifiers, potting, packaging, and insulation. A distinction is made between core materials and core assemblies since core assemblies are usually complete packages supplied by a manufacturer to deliver specified performance for various environmental conditions such as pressure, shock, vibration, etc. Atmosphere denotes any media surrounding and influencing the amplifier. For example, oxygen can be considered an adverse atmosphere in the same sense that humidity is considered adverse if a core is damaged at high temperatures due to oxidation. The chart is subject to change as the need arises.

This first report of the Working Group is a progress report and its purpose: to

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted July 11, 1960; made available for printing October 21, 1960.

Personnel of the Working Group on Magnetic Amplifier Materials Under Extreme Environmental Conditions of the AIEE Materials Subcommittee of the Magnetic Amplifier Committee, are F. GOURASH, *Chairman;* A. A. Annunziato, R. E. Anderson, J. S. Hamill, C. A. Harper, I. I. Kaplan, D. A. Kapp, R. S. Key, V. E. Legg, J. E. Mitch, W. T. Mitchell, R. A. Parnell, and W. Schlosser.

ENVIRONMENT		TEMPERATURE	NUCLEAR RADIATION	MECHANICAL STRESS	ATMOSPHERE
ABBREVIATION		TEMP	NUCL	MECH	ATM
COMPONENT	ABBREVIATION				
CORE MATERIALS	CORE MAT				
CORE ASSEMBLIES	CORE ASS'Y				
WIRE CONDUCTOR	WIRE				
RECTIFIERS	RECT				
POTTING, PACKAGING, INSULATION	PKG				

Fig. 1. Proposed environments and magnetic amplifier components

acquaint those interested with our activities and goals. Comments and suggestions are invited, particularly for the areas which are incompletely covered by the work to date. The first row of blocks of the chart concerning the effects of the listed environments on the properties of magnetic materials used in magnetic amplifier cores is described in this report. Since the main source of information is the literature available, a complete bibliography of all material on environmental conditions considered will be published with each report. Subsequent reports will be made on additional blocks in the chart as the investigations progress.

Core Materials

The magnetic core is one of the most essential components of the magnetic amplifier. The behavior of the magnetic materials important for magnetic amplifiers is detailed here. The term "extreme environmental conditions" implies conditions radically different from those for which magnetic amplifier designs are already adequate. The range covered by the word extreme extends from the conventional conditions to limits beyond which magnetic materials become nonoperative, or beyond which it is otherwise impracticable to operate.

The constant-current flux-reset test data as outlined in AIEE "Test Procedure for Toroidal Magnetic Amplifier Cores," Standard no. *432*, would be desirable in describing the effects of extreme environments on the behavior of cores for magnetic amplifiers. Unfortunately this type of data is unavailable but d-c test data which describe basic properties are available. The d-c data can be used to screen the various materials for operation under specific environments.

The basic a-c magnetic properties considered important in amplifier design are: saturation induction B_s, coercive force H_c, residual induction B_r, squareness ratio B_r/B_s, maximum permeability μ_m, initial permeability μ_o, insofar as these latter provide convenient and delicate gages of the sensitivity of a material to environmental influences. Environmental conditions which may tend toward extremes are temperature, ranging up to 500 C (degrees centigrade) or more, hydrostatic pressure ranging up to 20,000 psi (pounds per square inch), irradiation such as that near a nuclear reactor, and possibly some combinations of these.

The report is limited in that it considers only those materials which are available commercially to magnetic amplifier designers. Data on less common materials are available in the literature, but it remains for manufacturers to make them publicly available, particularly if such materials are deemed superior to current materials. Commercially available materials considered to be of importance in magnetic amplifiers include those in Table I.

Temperature Effects

The effects of temperature on magnetic properties are not great over the room-temperature range nor over the common extremes of weather, arctic to tropical sunshine. There are considerable effects when temperatures are increased still higher. Inevitably, the Curie temperature marks the point above which ferromagnetic properties disappear completely. Serious declines of properties can occur at much lower temperatures than the Curie temperature. These impairments are most marked with those materials which owe their superior magnetic properties to special heat treatments. The functions of heat treatment include relieving strains, purification, or, in the case of heat treating in a magnetic field, orienting domains. The heat-treating temperatures for magnetic materials usually exceed or are in the vicinity of the Curie temperatures. It may be reasonable to expect that raising the temperature of some materials, particularly those with lower Curie temperatures, would effectively give a further heat treatment which may well erase the benefits of prior favorable heat treatments. Such materials include the oriented alloys, and supermalloy and supermendur. The least sensitive materials are Mo-permalloy and nonoriented silicon steel, but they are also somewhat less desirable for moderate temperature applications.

Data for the magnetic materials are given in Table II. The data at higher temperatures are expressed as percentage changes from room-temperature data in order to reveal the trend of magnetic properties which may be expected to apply to similar materials, regardless of some deviations in room-temperature data.

Pressure Effects

Mechanical stress on magnetic materials is recognized as very harmful to magnetic properties when it is applied in any one direction. Such stress is encountered in applying windings on cores, in handling and in mounting them.

Omnidirectional stresses, however, are found to be relatively harmless to magnetic properties. That is, metallic

Table I. Magnetic Materials for Magnetic Amplifier Cores

Generic Name	Approximate Composition, %	American Trade Names*
Nonoriented silicon steel	4Si, 96Fe(iron)	Transformer A, core iron C: Trancor *M-14*
Oriented silicon steel	3Si, 97Fe	Silectron, Trancor, Oriented T, Hipersil
Oriented 50 Ni(nickel)Fe	50Ni, 50Fe	Deltamax, Hy-Ra *49*, Square-Mu, Orthonol, Hipernik *V*
Mo-permalloy	4Mo(molybdenum), 79Ni, 17Fe	4-79 Mo-permalloy, HyMu *80*, Hipernom
Mumetal	2Cr(chromium), 5Cu(copper), 77Ni, 16Fe	Mumetal
Supermalloy	5Mo, 79Ni, 16Fe	Supermalloy
Supermendur	2V(vanadium), 49Co(cobalt), 49Fe	Supermendur
Ferrites	Various	

* Typical but incomplete list. New names appear whenever inspired by sales or special technical appeal.

Table II. Effects of Temperature on Magnetic Cores in Air[1]

Material	Temperature, C	μ_m	H_c	$B_r - K_g$	$B_s - K_s$	B_r/B_s, %
Nonoriented	20	13,000	0.40	10.2	15.0	69
Si steel	200	+ 8%	− 5%		− 1%	
Curie temperature 690 C	400	+22%	− 27%		− 4%	
Transformer A	600	+46%	− 62%	−33%	−33%	− 3%
Return	20		+ 5%	0	0	0
Oriented	20	25,000	0.30	14.5	18.5	78
Si steel	200	− 8%	− 20%	− 5%	− 1%	− 4%
Curie temperature 710 C	400	−12%	− 40%	−25%	−10%	−17%
Silectron	600		− 70%		−23%	−28%
Return	20					−35%
Oriented 50 Ni-Fe	20	88,000	0.52*	14.5	15.9	91
Curie temperature 440 C	200	−12%	− 27%	−14%	−16%	+ 3%
Orthonol	400	−26%	− 65%	−50%	−48%	− 3%
Return	20		− 19%			
Mo-permalloy	20	140,000	0.025	5.8	8.6	68
Curie temperature 460 C	200	− 7%	− 16%	−33%	−13%	−23%
	400	−58%	− 24%	−73%	−53%	−46%
Return	20	Not permanently affected				
Supermendur	20	60,000	0.28	21.0	24.0	88
Curie temperature 980 C	200	+13%	− 7%	− 7%	− 3%	− 6%
	400	−62%	+ 4%	−32%	− 6%	−24%
	600	−60%	− 21%	−12%	−12%	0
	700	−53%	+168%	−81%	−31%	−74%
Return	20	Superior properties destroyed				

Time interval less than 4 hours.

* Values obtained from 60-cps tests.

Table III. Effect of Hydraulic Pressure to 20,000 Psi on Magnetic Materials

Material	$H_c - O_e$	$B_r - K_g$	Change at 20,000 Psi, %	μ_0	Change at 20,000 Psi, %	Change at Return, %
Oriented Si steel	0.29	16.0	0	700	+ 3	+4
Mo-permalloy	0.015	5.4	0	9,250	− 1.5	+5
Supermalloy	0.009	4.2		52,800	+ 0.4	+0.8
Supermendur	0.28	22.5				
Ferrites:						
Ferramic Q_1				145	+ 3.5	0
Manganese-zinc ferrite				2,200	−54	−1

cores, suspended in a liquid medium, are not markedly affected by pressures as high as encountered in the ocean depths. This has led to the mounting of tape cores in conformed core boxes, which are filled with a more or less fluid medium to suspend the cores hydrostatically. Effects of pressure on samples immersed in oil are shown in Table III.

Radiation Effects

Prolonged radiation, such as that in the immediate proximity of an active reactor, affects magnetic properties much the same as prolonged heating. Those materials which owe their distinctive properties to special heat treatments are most rapidly and permanently affected by high-energy radiation. Materials such as ferrites, which have low Curie temperatures, are impaired magnetically if their temperature rises excessively, due either to proximity to a heat source or to internal conversion of radiant energy into heat. Otherwise, ferrites are notably immune to radiation damage and to either temporary or long-time exposure. Table IV shows the effect of integrated fast flux (nvt) of 10^{17} neutrons per centimeter[2].

Atmosphere

The environment "atmosphere" listed in Fig. 1 denotes any media surrounding and influencing the amplifier. For example, it was noted that at the extreme high temperature of approximately 500 C, the core may be damaged through oxidation from the core surroundings. Thus, in this sense, oxygen is considered an adverse atmosphere just as the more ordinary items, such as humidity, are known as adverse.

Oxidation at high temperatures has the effects of destroying the interlaminar resistance and producing mechanical bloating or dimensional changes that cause strains in the cores. The material least liable to oxidize is Mo-permalloy; however, none of the materials should be exposed to strong oxidizing conditions at elevated temperatures. Supermendur cores exhibited anomalous transient effects when tested in an air atmosphere at 500 C, and upon return to room temperature the cores exhibited permanent magnetic degradation. These effects were attributed to oxidation as a result of the high-temperature cycling.

Table IV. Effect of Nuclear Radiation $nvt = 10^{17}$ on Magnetic Materials in Per-Cent Change from Original Properties[11]

Material	μ_0	$H_c - O_e$	Ratio B_r/B_m
Oriented 50 Ni-Fe	2,400	0.33	0.84
Change	2%	+ 12%	− 9%
Return	21%	+ 25%	− 6%
Mo-Permalloy	20,000	0.051	0.78
Change	84%	+150%	−45%
Return	84%	+200%	−47%
Supermalloy	48,000	0.035	
Change	96%	+300%	
Return	96%	+320%	
Supermendur	800	0.68	0.82
Change	2%	+ 40%	+ 2%
Return	17%	+ 60%	+ 8%
Ferrites:			
Ferramic S-6	1,100		0.75
Change	+15%		+ 2%
Return	0%		+ 2%

References

References on Temperature Effects

1. EFFECTS OF ULTRAHIGH TEMPERATURE ON MAGNETIC PROPERTIES OF CORE MATERIALS, **Michael Pasnak, Richard Lundsten.** *AIEE Transactions,* pt. I (*Communication and Electronics*), vol. 78, 1959 (Jan. 1960 section), pp. 1033–39.

2. TEMPERATURE DEPENDENCE OF MAGNETIC PROPERTIES OF SI-IRON, **C. W. Chen.** *Journal of Applied Physics,* New York, N. Y., vol. 29, 1958, p. 1337.

3. AN EXPERIMENTAL STUDY OF MAGNETIC MATERIALS FOR USE IN ULTRAHIGH-TEMPERATURE ELECTRONIC TRANSFORMERS, **H. B. Harms.** *AIEE Transactions,* pt. I (*Communication and Electronics*), vol. 77, May 1958, pp. 181–84.

4. FEASIBILITY OF USING MAGNETIC AMPLIFIERS WITH 2V-PERMENDUR CORES AT TEMPERATURES UP TO 500°C, **W. H. Raskin.** *USNOL Report 4390,* U. S. Naval Ordnance Laboratory, Silver Spring, Md., 1956.

5. PERFORMANCE OF PERMANENT MAGNETS AT ELEVATED TEMPERATURES, **W. H. Roberts.** *Journal of Applied Physics,* vol. 29, 1958, p. 405.

6. EFFECTS OF TEMPERATURE ON MAGNETIC PROPERTIES OF NICKEL-IRON ALLOYS, **J. F. Fritz, J. J. Clark.** *Electrical Manufacturing,* New York, N. Y., Nov. 1958.

7. EFFECTS OF TEMPERATURE ON A-C MAGNETIC PROPERTIES OF NICKEL-IRON ALLOYS, **J. F. Fritz, J. J. Clark.** *Ibid.,* Dec. 1959.

References on Pressure and Mechanical Effects

8. EFFECTS OF HYDROSTATIC PRESSURE ON PROPERTIES OF MAGNETIC MATERIALS, **R. E. Alley, V. E. Legg.** *Journal of Applied Physics,* vol. 31, Mar. 1960.

9. POLYMORPHISM IN IRON AT HIGH PRESSURE, **D. Bancroft, E. L. Peterson.** *Ibid.,* vol. 27, 1956, p. 291.

10. EFFECTS OF TRANSVERSE COMPRESSIONAL STRESS ON MAGNETIC LAMINATIONS, **R. E. Fischell.** *AIEE Transactions,* pt. I (*Communication and Electronics*), vol. 74, May 1955, pp. 190–93.

11. IN-PILE MEASUREMENTS OF RADIATION EFFECTS IN MAGNETIC MATERIALS, **R. E. Alley.** *Journal of Applied Physics,* vol. 30, 1959, p. 284S.

Reference on Atmosphere Effects

12. CHARACTERISTICS OF SUPERMENDUR AT 500°C, **M. Lauriente, G. E. Lynn.** *Journal of Applied Physics,* vol. 31, no. 5, May 1960, supplement, p. 237S.

Effective Value of Direct-Voltage Ripple

E. J. DIEBOLD
MEMBER AIEE

Synopsis: Rectifiers produce a ripple of the direct-voltage output. Because it has many different shapes, it is hard to describe in simple *terms* except by the effective (root-mean-square) value of the ripple voltage, giving only one number to designate its magnitude. Within the limitations given in the paper, all ripples depend only on voltage magnitude, repetition rate, overlap (time of current commutation), and phase delay angles. Calculation of the magnitude is attempted by various methods; the most straightforward one uses two normalized functions of overlap and repetition rate (expressed as per-unit values). Figures in the paper give these functions within the range of validity and examples show how to use the method and interpret the results. Without phase control the resulting relative ripple magnitude (ripple voltage over direct voltage) is plotted against both repetition rate and overlap.

SEMICONDUCTOR devices are changing the field of rectification by opening new possibilities of rectifier application. Replacement of existing rectifier equipment presents few problems, but entirely new fields of replacement of rotary equipment by rectifiers are more difficult to assess. Evaluation of new approaches requires an accurate re-examination of old problems, among which the magnitude of the ripple, the cost of its elimination, and the cost of the lost power due to ripple are important.

The objective of this paper is to present a simple method of comparing the magnitude of direct-voltage ripples and to approximate their effect on a practical d-c load. Knowing the ripple magnitude, both as current and voltage, can be very useful, particularly when essentially different types of equipment (rotating versus rectifier equipment of various types) are compared. Competing claims by different manufacturers may be evaluated, including a method to assess variations in operating costs (a-c power lost in the d-c load) against capital costs (phase shifters and interphase transformers).

Direct-voltage ripples have several effects which may be of economic importance:

1. Ripple noise in the d-c load, e.g., telephone station battery chargers.
2. Telephone interference of d-c lines (traction).
3. Inductive heating of the d-c system (utilities).
4. A-c losses in the d-c load.
5. Reflected ripple currents in the a-c line causing telephone interference and poor distortion power factor.

Regarding the a-c losses in the d-c load, for heating applications this power dissipation is not necessarily a loss. For electrochemical applications it represents a loss in production. The power loss depends on the ripple voltage magnitude and frequency as they appear on the load (not at the individual rectifier terminals) and also on the electrical characteristics of the load. Notably, if many equal but noninterconnected d-c loads are present in the same plant, the plant distortion factor may be unity and the telephone interference factor may be negligible; yet the a-c losses in the d-c load may be very large. In the usual large electrochemical plants, the ripple frequency is much higher and the ripple magnitude is much less in the load than in the individual rectifiers; hence there is no problem of power loss. In other installations this may be entirely different, e.g., in some utility rectifiers with 3-wire connection wherein a 6-phase ripple appears between lines, and a 3-phase ripple appears between line and neutral.

A problem of economic interest is the comparison of existing equipment having voltage control by phase delay with new equipment without phase delay. Generally the means of providing voltage control without phase delay are more expensive but the difference in ripple magnitude, if exactly known, may be an important factor in deciding between them. Economically important evaluations require technical information in a form which can be compared easily, obtained quickly, and changed without too much difficulty if the conditions are changed. Expediency and simplicity are more important than accuracy because many economic considerations are not made if they are too cumbersome or if the results of the evaluation are not directly comparable. For this reason it was decided that the effective value of the direct-voltage ripple is the most useful single term describing the ripple of any d-c waveshape; this value is the root mean square of the voltage difference between the instantaneous and the average direct voltage. Any waveshape can be assigned a ripple number, although the differences in shape and frequency may be very substantial. Having one characteristic number allows certain conclusions to be drawn rigidly, and others inaccurately (though these are still useful), whereas some further evaluations require a very lengthy calculation. By expressing the effective value of the direct-voltage ripple in per cent of direct voltage, a normalized figure is obtained which can be compared to figures such as efficiency or per-unit losses.

Efficiency

Rectifier equipment with high-quality rectifier devices is usually purchased on economic merits derived from the high-power efficiency. If this equipment produces a high amount of alternating voltage on the d-c terminals, causing a high alternating current and a-c power loss in the d-c load, the efficiency which was the reason for purchasing the equipment becomes more of a vendor's sales gimmick than an economic reality for the buyer. Manufacturers of equipment state the efficiency as power efficiency (output watts over input watts), whereas users are interested in conversion efficiency (output volt-amperes over input watts). For infinitely inductive load or zero ripple output the two efficiencies are equal. Published data on conversion efficiency are for purely resistive load. Capacitive or counterelectromotive loads have a much lower conversion efficiency than these published values. Electrochemical loads are quite complex; they are capacitive, counterelectromotive, resistive, and inductive all at once; therefore the conversion efficiency may be higher or lower than the resistive value. Battery loads and capacitor banks are notorious for poor conversion efficiency, which is frequently compounded by additional losses in the rectifier equipment. If the rectifier output power is resold, the question of conversion versus power efficiency depends on the metering principle used. If a true kw-hour meter is used, only the power efficiency is important.

In economic evaluation by any user, first the decision must be made as to which efficiency is really important. If

---

Paper 60-211, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted November 27, 1959; made available for printing September 21, 1960.

E. J. Diebold is with the International Rectifier Corporation, El Segundo, Calif.

I. K. Dortort and D. W. Borst contributed valuable advice and references to this paper.

it is the conversion efficiency, the rectifier losses can usually be found accurately enough from the power efficiency of the rectifier equipment. A-c losses in the d-c load may be determined approximately by the method outlined in this paper.

Nomenclature

C = capacity of capacitive load, farads
Δe = instantaneous ripple voltage
ΔE = rms value ripple voltage
E = alternating voltage, rms value $=(1/\sqrt{2})E_0$
E_0 = crest value of transformer secondary voltage, measured in such a way that it is also equal to crest value of direct voltage
E_0' = apparent crest voltage, long overlap
e_d = instantaneous direct voltage, at angle θ
E_d = average value of direct voltage
E_e = rms value of direct voltage
E_r = average voltage appearing on load resistance
i_d = instantaneous direct current, amp (amperes)
I_d = average value of direct current, amp
Δi = instantaneous ripple current, amp
ΔI = rms value ripple current, amp
K_1, K_2 = normalized numerical functions of ripple frequency and overlap
L = inductance of load, henrys
p = number of secondary (commutating) phases in each group
P_d = true d-c power in load, watts
P_r = ripple power (a-c power dissipated in d-c load), watts
q = number of identical d-c ripples per cycle of a-c frequency
Q_1, Q_2, Q_3 = constants depending on q
R = resistance of load
T = inductive time constant of load
u = overlap angle
u' = apparent overlap angle for large overlap
U_1, U_2 = functions of ripple frequency and overlap
v = per-unit overlap $= uq/\pi$
V_1, V_2, V_3, V_4 = algebraic functions of v
W_1, W_2, W_3, W_4 = algebraic functions of v
X = reactance from anode to neutral of the commutating circuit, ohms
α = phase delay angle
α' = apparent phase delay angle for large overlap
$(\alpha+u)^*$ = maximum phase delay and overlap under assumption 5
$\cos\phi$ = displacement power factor of ripple voltage and current
θ = variable angle for integration
ω = frequency, radians per second

Assumptions

Theoretical analysis of a practical problem leads usually to a set of rigid assumptions. In this case, the assumptions are common to most theoretical papers on rectification. It is, however, important to be aware of these restrictions because when they are not met, the theory is wrong.

1. The direct current flows during the entire cycle without interruption and is commutated from phase to phase. This eliminates the single-phase half-wave rectifier and all rectifier circuits, such as magnetic amplifiers, in which it is expected that the direct current returns to zero before the next ripple can commence. If in a magnetic amplifier the continuous flow of direct current is assured by a free-wheeling diode across the output of the rectifier, this theory is still not valid, except when the diode does not carry current.

2. It is assumed that each phase of the rectifier carries the same amount of current and has the same voltage waveshape at the same time intervals. This is equivalent to a perfectly balanced polyphase a-c system. Whenever it is attempted to reduce the direct-voltage ripple to a very small magnitude by going to a very large number of phases, the fact that this second assumption is not met leads to false results. The user of this theory should have reliable figures on the maximum phase unbalance from the a-c systems. The phase unbalance will be rectified as is a single-phase alternating voltage, no matter how many phases are used in the rectifier. Reduction of the ripple can only be achieved by eliminating the phase unbalance, or filtering out the ripple.

3. Voltage control is effected by phase delay or variable transformer only. This excludes variable impedance devices such as primary saturable reactors. The self-saturating type of magnetic amplifiers are acceptable under this assumption as long as assumption 1 is met.

4. The overlap is shorter than the duration of one full ripple. This assumption is not quite rigid; it can be bypassed under certain conditions with an approach that leads to serviceable, although not quite correct, results as indicated in Appendix I.

5. The direct-voltage output is the sum of one or more groups of rectifiers, each group consisting of a number of rectifier elements which form a complete commutating cycle, for example, the 3-phase double-Y rectifier contains two groups of three elements or phases. It is assumed that the current overlap of each group occurs simultaneously only between two phases. Numerical values of overlap limits for various numbers of phases per group are represented in Appendix I.

Method of Approach

The classical approach to determine the effective value of direct-voltage ripples is by Fourier analysis. The direct voltage waveshape (determined by the number of ripples per fundamental a-c wave, phase delay, and overlap) can be drawn up as a succession of short pieces of sine waves. By using the Fourier integrals, the coefficients of the Fourier series may be determined. The coefficients are amplitudes of harmonic voltages, and the sum of their squares gives the square of the effective value of the ripple voltage. There are two coefficients for each harmonic, except for the fundamental. In most practical cases the series of squares of harmonic amplitudes converges quite rapidly.[1–6]

The method outlined above is accurate; generally valid equations for the individual coefficients of the Fourier series which are not too complex have been published. If a general equation is wanted, containing the sum of the square of the coefficients, totalized from the first significant harmonic to infinity, the equation becomes impractical. Thus the method requires calculation of every problem from the beginning, through a rather elaborate process, and repetition of the same multitude of calculations with different numbers.

The author tried a second method of approach wherein the voltage waveshape was expressed as a succession of pieces of sine waves; these expressions were squared and integrated and the result is a single, but unwieldy, expression. The approach seems to be more straightforward and simpler than the approach by Fourier analysis. Any attempt to use the equations gained by the second approach leads to very frustrating numerical work. The expressions resolve to differences of numbers which are almost equal. Even calculation of great accuracy, with the use of tables of trigonometric functions with seven decimals and mechanical calculators, produces results which are too inaccurate to be useful or they may be completely wrong due to a small error. Evaluation, rapid judgment, or quick comparison cannot be made by this method.

It was necessary to find a third approach which would consist of taking the results of the foregoing and pairing all the functions which are approximately equal but subtracted from each other, then of searching for new functions which express these differences as simply as possible. A better expression might have been found, except for some unwieldy constants which had to be calculated once and for all. The calculation of these constants, containing the difficulties of numerical work, led to a further understanding of the problem, which suggested a fourth method of approach.

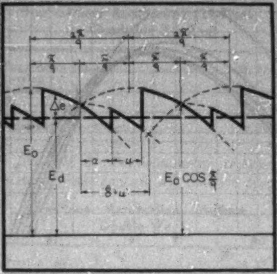


Fig. 1. Direct-voltage waveshape

The final method derives from expressing, as series, the constants obtained from the difference of almost equal functions. With this approach the small differences of large numbers disappear completely. The final numerical work is greatly simplified by the use of two functions which depend only on the number of ripples per a-c cycles and the overlap; these parameters are given graphically.

Effective Value of Direct Voltage

According to the definitions of the nomenclature, the direct-voltage ripple, instantaneous value, is given by equation 1. The effective value (rms) is given by equation 2.

$$\Delta e = e_d - E_d \tag{1}$$

$$\Delta E^2 = q/2\pi \int_0^{2\pi/q} \Delta e_d^2 d\theta = E_e^2 - E_d^2 \tag{2}$$

The effective value of the direct voltage is obtained by integrating the voltage-time wave of Fig. 1 as in equation 3:

$$E_e^2 = q/2\pi E_0^2 \left(\int_{-\pi/q+\alpha+u}^{\pi/q+\alpha} \cos^2 \theta d\theta + \cos^2 \pi/q \int_{\alpha}^{\alpha+u} \cos^2 \theta d\theta \right) \tag{3}$$

After performing the integration we obtain

$$E_e^2 = \left(\frac{E_0 \sin \pi/q}{2\pi/q} \right)^2 \left\{ \frac{\pi}{q} \cos(2\alpha + u) \left[2 \operatorname{ctg} \frac{\pi}{q} \cos u + \sin u \right] + 2 \left(\frac{\pi/q}{\sin \pi/q} \right)^2 - \frac{\pi}{q} u \right\} \tag{4}$$

The most direct approach would be to calculate the effective value of the direct voltage according to equation 4, the average value from equation 40 or 41 and, by inserting in equation 2, obtain the effective value of the direct-voltage ripple. Because of the very small differences between the square of the effective value and the square of the average value this method is not recommended.

From equation 41, we form

$$E_d^2 = \left(\frac{E_0 \sin \pi/q}{2\pi/q} \right)^2 [\cos(2\alpha + u) + 1](\cos u + 1) \tag{5}$$

The similarity between equations 4 and 5 is deliberate, for forming the difference which is obtained by inserting equations 4 and 5 in equation 2, to form

$$\Delta E^2 = \left(\frac{E_0 \sin \pi/q}{2\pi/q} \right)^2 \left\{ 2 \left[\left(\frac{\pi/q}{\sin \pi/q} \right)^2 - 1 \right] - \frac{\pi}{q} u + (1 - \cos u) - \cos(2\alpha + u) \left[\cos u \left(1 - \frac{2\pi}{q} ctg \frac{\pi}{q} \right) + 1 - \frac{\pi}{q} \sin u \right] \right\} \tag{6}$$

Equation 6 contains all the necessary elements for calculating the effective value of the direct-voltage ripple. It can be simplified at the risk of being even more impractical than it already is. First, it contains many trigonometric functions which must be taken from tables which always have a limited accuracy. The value in the brackets, for most practical applications, is in the order of magnitude of 0.001 to 0.000001, whereas the terms themselves are in the vicinity of unity. It appears from the equation that the first half of the bracketed term is independent of the phase delay and the second half contains the phase delay only once.

Elimination of Differences

The approach to simplifying equation 6 was possible only through a major reorganization, as shown in equation 7. The functions U_1 and U_2 are given by equations 8 and 9, containing the functions Q_1 and Q_2 and Q_3, shown in equations 10 through 12 with the interrelation equation 13.

$$\Delta E^2 = \left(\frac{E_0 \sin \pi/q}{2\pi/q} \right)^2 \left[U_1 2 \sin^2 \left(\alpha + \frac{u}{2} \right) + U_2 \cos(2\alpha + u) \right] \tag{7}$$

$$U_1 = Q_1 - \frac{\pi}{q} u + (1 - \cos u) \tag{8}$$

$$U_2 = Q_2 - \frac{\pi}{q}(u - \sin u) + 2Q_3(1 - \cos u) \tag{9}$$

$$Q_1 = 2 \left[\left(\frac{\pi/q}{\sin \pi/q} \right)^2 - 1 \right] \tag{10}$$

$$Q_2 = 2 \left[\left(\frac{\pi/q}{\sin \pi/q} \right)^2 + \frac{\pi}{q} \operatorname{ctg} \frac{\pi}{q} - 2 \right] \tag{11}$$

$$Q_3 = 1 - \frac{\pi}{q} \operatorname{ctg} \frac{\pi}{q} \tag{12}$$

$$Q_2 = Q_1 - 2Q_3 \tag{13}$$

The reorganization of equation 6 into the form of equations 7 through 13 has the advantage that equation 7 contains only additive terms and that these terms are easily computed if the functions U_1 and U_2 are known. These are functions of only the overlap u and the number q of the d-c ripples per a-c cycles; they therefore lend themselves to being plotted in a graph because the variable q is always an integer and the overlap u is always limited, due to assumption 4. The functions Q_1, Q_2, and Q_3 depend only on the variable q.

The numerical values given later were determined by means of equations 7 through 13 for $q=2$, $q=3$, and $q=4$. For higher values of q, even these equations are unsatisfactory. The difficulty is again the formation of differences of almost equal terms in equations 10 through 13, and subsequently in equations 8 and 9. This can be overcome by extremely accurate calculation but is very conducive to errors.

Development Into Series

Equations 8 through 12 invite development into series, to eliminate those terms which are directly subtracted. Thus equation 10 resolves into equation 14, equation 11 into equation 15, and equation 12 into equation 16.

$$Q_1 = \frac{2}{3} \left(\frac{\pi}{q} \right)^2 \left[1 + \frac{1}{5} \left(\frac{\pi}{q} \right)^2 + \frac{2}{63} \left(\frac{\pi}{q} \right)^4 + \frac{1}{225} \left(\frac{\pi}{q} \right)^6 + \frac{2}{3{,}465} \left(\frac{\pi}{q} \right)^8 + \ldots \right] \tag{14}$$

$$Q_2 = \frac{4}{45} \left(\frac{\pi}{q} \right)^4 \left[1 + \frac{4}{21} \left(\frac{\pi}{q} \right)^2 + \frac{1}{35} \left(\frac{\pi}{q} \right)^4 + \frac{4}{693} \left(\frac{\pi}{q} \right)^6 + \ldots \right] \tag{15}$$

$$Q_3 = \frac{1}{3} \left(\frac{\pi}{q} \right)^2 \left[1 + \frac{1}{15} \left(\frac{\pi}{q} \right)^2 + \frac{2}{315} \left(\frac{\pi}{q} \right)^4 + \frac{1}{1{,}575} \left(\frac{\pi}{q} \right)^6 + \ldots \right] \tag{16}$$

These equations show why the numerical approach by trigonometric functions is so unsatisfactory, because either the first two or the first four terms disappear completely. Developing equation 8 in series, by using equation 14, gives equations 17 through 22. In these the variable v is the per-unit overlap (compared to the duration a d-c ripple). They show that the value of U_1 is very similar to the value of Q_1 except that in the former the coefficients V_1, V_2, etc., are

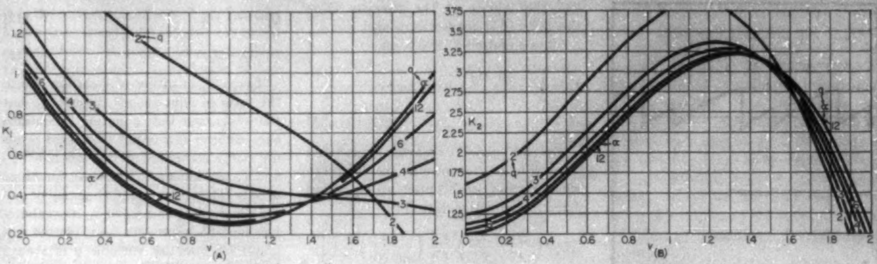


Fig. 2. Normalized functions for several ripple numbers q, plotted against relative overlap v

A—Function K_1 B—Function K_2

different from unity. Thus, we have finally reduced the variable U_1 to a series, having constant coefficients in the variable q, and variable coefficients with the variable v, given by simple algebraic equations 19 through 22. The function U_1 is proportional to a new variable K_1, which is close to unity.

$$U_1=2/3\left(\frac{\pi}{q}\right)^2 K_1 \tag{17}$$

$$K_1=V_1+\frac{1}{5}\left(\frac{\pi}{q}\right)^2 V_2+\frac{2}{63}\left(\frac{\pi}{q}\right)^4 V_3+\frac{1}{225}\left(\frac{\pi}{q}\right)^6 V_4+\dots \tag{18}$$

$$V_1=1-\frac{3}{2}v+\frac{3}{4}v^2 \tag{19}$$

$$V_2=1-\frac{5}{16}v^4 \tag{20}$$

$$V_3=1+\frac{21}{320}v^6 \tag{21}$$

$$V_4=1-\frac{15}{1{,}792}v^8 \tag{22}$$

A similar series of equations 23 through 28 can be written for the function U_2, which depends on a function K_2. K_2 is near to unity, having again coefficients of q similar to equation 15 and functions of v given by W_1, etc., in equations 25 through 28.

$$U_2=4/45\left(\frac{\pi}{q}\right)^2 K_2 \tag{23}$$

$$K_2=W_1+\frac{4}{21}\left(\frac{\pi}{q}\right)^2 W_2+\frac{1}{35}\left(\frac{\pi}{q}\right)^4 W_3+\frac{4}{693}\left(\frac{\pi}{q}\right)^6 W_4+\dots \tag{24}$$

$$W_1=1+\frac{15}{4}v^2\left(1-\frac{v}{2}\right) \tag{25}$$

$$W_2=1+\frac{21}{16}v^2\left(1-\frac{5}{4}v^2+\frac{3}{8}v^3\right) \tag{26}$$

$$W_3=1+\frac{5}{6}v^2\left(1-\frac{7}{8}v^2+\frac{7}{16}v^4-\frac{3}{32}v^5\right) \tag{27}$$

$$W_4=1+\frac{33}{80}v^2\left(1-\frac{5}{6}v^2+\frac{7}{24}v^4-\frac{5}{64}v^6+\frac{5}{384}v^7\right) \tag{28}$$

From this it is apparent that the functions K_1 and K_2 can be plotted as variables of the number of d-c ripples q and the per-unit overlap v, without obtaining an excessive spread of the different values due to the q-depending terms in equations 17 and 23. Values of functions K_1 and K_2 are plotted in Fig. 2.

When the per-unit functions shown in Fig. 2 are used, the variations with the number q of d-c ripples per a-c cycles are very striking. With over six d-c ripples per a-c cycle, the variations of the functions K_1 and K_2 are small, almost without deviation between 12 and an infinite number of d-c ripples per cycle. Thus the calculation is quite simple for a high number of d-c ripples, in exact opposition to the problem of calculation by the original method. On the other hand, equations 17 and 23 indicate a strong dependancy of the functions U_1 and U_2 on the number of phases q, but in a very simple form.

General Equation

By combining equations 7, 41, 17, and 23, we finally obtain

$$\frac{\Delta E}{E_d}=\frac{\pi}{q}\frac{\sqrt{K_1\sin^2(\alpha+u/2)+\frac{1}{15}(\pi/q)^2K_2\cos(2\alpha+u)}}{\sqrt{3}\cos(\alpha+u/2)\cos u/2} \tag{29}$$

This gives the effective value of the direct-voltage ripple proportional to the average direct voltage. The functions K_1 and K_2 are shown in Fig. 2. With these, any value of relative direct-voltage ripple can be calculated rather quickly, and without using very accurate calculation or trigonometric tables with many significant numbers.

It would be desirable to plot the relative value of the d-c ripple for any number of d-c ripples q, phase delay angle α and overlap angle u (or relative overlap angle v). Since this involves three independent variables, the graphic representation would be difficult and the numerical work quite substantial. The numerical work might be undertaken by an electronic computer, at least for equations 17 through 28, whereas equation 29 would have to be rewritten for easier application in a computer.

Equation 29 with Fig. 2 provides a simple and sufficiently accurate method to calculate the magnitude of any direct-voltage ripple which fits the assumptions of this paper. For anyone who prefers to go beyond the scope of the paper, the equations have been simplified sufficiently to calculate a great number of values by a computer.

Examples

By the method described, a few examples have been calculated:

Example 1: Mechanical rectifier, 4% magnetic phase control, 10% mechanical phase delay, 6% reactive drop.

Six-Phase Rectification (6-Phase, Double-Way)

Phase delay angle	30.7°
Overlap	11.6°
Relative overlap	0.387
K_1	0.593
K_2	1.510
Relative ripple voltage	17.67%

Example 2: Mercury arc rectifier, 10% (grid control) phase delay, 6% reactive drop.

Twelve-Phase Rectification (12-Phase, Quadruple Zig-Zag)

Phase delay angle	25.8°
Overlap	12.9°
Relative overlap	0.860
K_1	0.262
K_2	2.59
Relative ripple voltage	4.94%

Example 3: Semiconductor rectifier, without phase control, 4% reactive drop.

Six-Phase Rectification (6-Phase, Double-Way)

Overlap	23.1°
Relative ripple voltage	7.28%

These examples illustrate the advantages of phase multiplication and operation without phase control.

Rectifiers Without Phase Delay

From equations 41 and 6, with equations 15 and 16, for zero phase delay, we obtain

$$\left(\frac{\Delta E}{E_d}\right)^2 = \frac{Q_2+2Q_1\sin^2 u-\frac{\pi}{q}\left(u-\frac{\sin 2u}{2}\right)+(1-\cos u)^2}{4\cos^4 u/2} \tag{30}$$

This equation could be developed into series for further simplification, but numerical calculation shows that up to $q=12$ the calculation is not too difficult or inadequate. Fig. 3(A), showing the result of the calculation, indicates the rapid reduction of the ripple as the number of d-c ripples per a-c cycle increases, and also the substantial influence of the overlap angle.

For $q=12$ and $q=24$ the extension of the range of assumption 4 was made in accordance with Appendix I. Since the error introduced by this simplification is only a few per cent, the curves are still of great interest. As an example: For the same relative overlap, the relative ripple decreases inversely, proportional to the square of the ripple number q, or slightly faster. For the same absolute overlap, e.g., 30 degrees, the decrease is much less, i.e., approximately inversely proportional. Furthermore, per-cent ripple values taken without overlap are quite meaningless for practical rectifiers, which always have a large overlap (20 to 30 degrees).

A-C Losses in D-C Load

Ripple magnitude and frequency, together with load characteristics, determine the ripple loss in the load. If the load is purely resistive (which is rarely the case) the ripple power is given by

$$\frac{P_r}{P_d}=\left(\frac{\Delta E}{E_d}\right)^2 \tag{31}$$

Thus a 10% ripple causes 1.0% power loss. With a counter-emf (electromotive force) and resistance in series (battery with its losses, electrochemical cell line) the ripple power loss is

$$\frac{P_r}{P_d}=\left(\frac{\Delta E}{E_d}\right)^2\left(\frac{E_d}{I_dR}\right) \tag{32}$$

In this example, with 10% ripple and 60% counter-emf ($I_dR=0.4\ E_d$) the power loss is 2.5%. Fortunately, real applications are much less severe, because of the d-c bus-bar inductance.

To consider the inductance (or also a capacitive load) correctly, the ripple should be subdivided into harmonic frequencies and amplitudes, and the power calculated for each frequency and totalized over all frequencies. This method is quite cumbersome and thus the entire subject is often avoided because of lack of an easier method.

Assuming that the ripple contains only the lowest significant harmonic, i.e., replacing the real shape of the ripple by a sine wave having the same fundamental frequency and effective (rms) value, a power loss may be calculated. It is easy to prove, for an inductive system, that the real power loss will be less than the one given by this assumption. For capacitive (noninductive) systems the real loss will be larger than the calculated value. A very large capacity (battery, capacitor bank, cell line) subjected to a relatively high ripple frequency (e.g., 360 cycles per second) may be approximated by a counter-emf with impedance added for leads, loop inductance, and internal resistance of the capacitor.

If ω is the angular frequency of the ripple, the per-unit ripple loss is given by

$$\frac{P_r}{P_d}=\left(\frac{\Delta E}{E_d}\right)^2\left(\frac{E_d}{I_dR}\right)\frac{1}{\left(1+\frac{\omega L}{R}\right)^2} \tag{33}$$

Without counter-emf (capacity) the term E_d/I_dR is unity. In the above example with 10% ripple, 60% counter-emf and $\omega L=R$, the ripple power is 1.25%.

Example 4: Rectifier, 14% phase control, 6% regulation, 6-phase calculated ripple voltage 17.7%. Load, 10,000-amp 500-volt total, 300-volt counter-electromotive force. Loop inductance 20 μh (microhenrys).

Thus $\omega L=0.0452$ ohm, $R=0.02$ ohm, and the ripple power over direct power is 1.28%. (It would be 7.83% without the loop inductance.) Most existing plants have high d-c loop inductance, which is

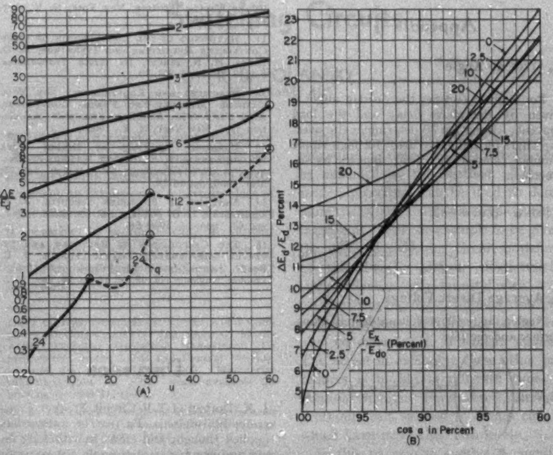


Fig. 3. Relative voltage ripple in per cent

A—For ripple numbers q, plotted against overlap u. Dashed lines are approximations; circled points designate limits of validity

B—For several values of commutating reactance plotted against phase control; valid for 6-phase rectification only very helpful. Still, compared with some calculations of high efficiency which consider tenths and hundredths of per cent, the above figure looks disagreeably high.

Example 5: Battery charger 48 volts, 1,200 amp, total internal resistance 8%, internal inductance of rectifier 8%, loop inductance of battery 6 μh.

Six-Phase Ripple, No Phase Control, 20 Degrees Overlap

Ripple voltage magnitude	6.8% (3.27 volts)
Total resistance (rectifier and battery)	3.2 milliohm
Total inductance (rectifier and battery)	14.5 μh
Ripple power neglecting inductance	5.8%
Ripple power considering inductance	0.0055%

Thus the internal inductance of the rectifier and the small loop inductance of the battery act as very good filters.

Without phase control and with a sufficient loop inductance (internal inductance of rectifier and of the load) the ripple loss in the d-c load is negligible.

Example 6: Six-phase rectification, expressing the phase delay in per cent of cos α, and the commutating reactance in per cent of E_x/E_{d0}; see equation 39. Results are plotted as per-cent rms ripple in Fig. 3(B).

Combining the effects of phase delay and commutating reactance results in a complex relationship. A singular condition exists at 6% to 7% phase delay in which a variation of commutating reactance from zero to 10% has no effect on the magnitude of the ripple. For lesser phase delay, the ripple increases with the commutating reactance. At greater phase delay, the ripple first decreases then increases with increasing commutating reactance.

Appendix I

Assumption 4

Beyond the range given in assumption 4 the direct-voltage waveshape is very similar to the one within the range, with the following exceptions; 1. The beginning and the ending of overlap have traded places. 2. The apparent crest voltage has been reduced by

$$E_0{}^1 = E_0\left(\frac{1+\cos\dfrac{2\pi}{q}}{2}\right) \qquad (34)$$

3. The apparent overlap voltage is less than the amount proportional to equation 34;

$$E_0 \cos \pi/q \cos 2\pi/q < E_0 \cos \pi/q \left(\frac{1+\cos 2\pi/q}{2}\right) \qquad (35)$$

Simplification

If we assume that the two sides of the inequality 35 are equal, the result will be slightly in error. The per-unit error will be commensurate with the per-unit difference of inequality 35.

The effective ripple can be calculated as within the range by setting

$$u' = u - 2\pi/q \qquad (36)$$

$$\alpha' = \alpha + \pi/q \qquad (37)$$

for $2\pi/q \leqq u \leqq 4\pi/q$ and multiplying the resulting ripple by

$$\Delta E' = \Delta E \frac{2}{(1+\cos 2\pi/q)} \qquad (38)$$

The per-unit ripple remains unchanged.

Example: Fig. 3(A) for $q=12$, shows an extension of the curve as dotted line determined by this method. $\pi/q = 15°$ $\alpha' = 15°$ $u' = u - 30°$

Assumption 5

Numerically, the maximum permissible phase delay plus overlap $(\alpha+u)^*$ is given by:

$$tg(\alpha+u)^* = \frac{\cos \pi/p - \cos 3\pi/p}{\sin 3\pi/p} = \frac{\sin 2\pi/p}{3/2 - 2\sin^2 \pi/p}$$

p, Phases per Group	$(\alpha+u)^*$, Degrees
3	90.0
4	63.5
6	40.9

Appendix II

I_a Overlap Angle

For a given rectifier the overlap angle is given by the phase delay and the load current. Equation 39 determines the overlap angle for inductive load (for high ripple number, almost any load appears as inductive).

$$\frac{1}{2}[\cos\alpha - \cos(\alpha+u)] = \frac{IX}{2E_0 \sin(\pi/p)} = \frac{E_x}{E_{d0}} \qquad (39)$$

I_b Average Direct Voltage

Neglecting resistive voltage drop in transformer and rectifier

$$E_d = E_0\left(\frac{\sin\dfrac{\pi}{q}}{\dfrac{\pi}{q}}\right)\frac{\cos\alpha + \cos(\alpha+u)}{2} \qquad (40)$$

$$E_d = E_0\left(\frac{\sin\dfrac{\pi}{q}}{\dfrac{\pi}{q}}\right)\cos(\alpha+u/2)\cos u/2 \qquad (41)$$

Equations 39 and 40 are equivalent.

I_c Phase Delay

For a given load and crest voltage, the phase delay determines the direct voltage by combining equations 39 and 40:

$$E_d = E_0\left(\frac{\sin\dfrac{\pi}{q}}{\dfrac{\pi}{q}}\right)\cos\alpha - \frac{IX\sin\dfrac{\pi}{q}}{2\dfrac{\pi}{q}\sin\dfrac{\pi}{p}} \qquad (42)$$

References

1. J. **Fischer Hinnen.** *Elektrotechnische Zeitschrift*, Wuppertal-Elberfeld, Germany, vol. 22, 1901, p. 396.

2. A SIMPLE GRAPHICAL METHOD FOR THE HARMONIC ANALYSIS OF A CYCLICAL FUNCTION, **J. R. Ashworth.** *Electrician*, London, England, vol. 67, 1911, p. 888.

3. THE ANALYSIS OF PERIODICAL WAVES, **L. W. Chubb.** *Electrical Journal*, London, England, vol. 10, 1914, p. 91.

4. DISCUSSION OF A GRAPHICAL FOURIER ANALYSIS, **Thomas C. Blow.** *Electronics*, New York, N. Y., Dec. 1947, pp. 194–98.

5. A NEW FOURIER COEFFICIENT HARMONIC ANALYZER, **S. Charp.** *AIEE Transactions*, vol. 68, pt. I, 1949, pp. 644–51.

6. PRINCIPLES OF MERCURY-ARC RECTIFIERS AND THEIR CIRCUITS (book), **D. C. Prince, F. B. Vogdes.** McGraw-Hill Book Company, Inc., New York, N. Y., 1927.

7. MERCURY-ARC POWER RECTIFIERS, THEORY AND PRACTICE (book), **O. K. Marti, H. Winograd.** McGraw-Hill Book Company, Inc., 1930.

8. OUTPUT WAVE SHAPE OF CONTROLLED RECTIFIERS, **F. O. Stebbins, C. W. Frick.** *AIEE Transactions (Electrical Engineering)*, vol. 53, Sept. 1934, 1259–65.

9. CURRENT AND VOLTAGE WAVE SHAPE OF MERCURY-ARC RECTIFIERS, **H. D. Brown, J. J. Smith.** *Ibid.*, vol. 52, Dec. 1933, pp. 973–86.

10. RECTIFIER WAVE SHAPE. *Publicction no. E1*, Edison Electric Institute, New York, N. Y., Apr. 1937.

11. HARMONICS IN THE A-C CIRCUITS OF GRID-CONTROLLED RECTIFIERS AND INVERTERS, **R. D. Evans, H. N. Muller, Jr.** *AIEE Transactions*, vol. 58, 1939, pp. 861–70.

12. PRACTICES AND REQUIREMENTS FOR POOL CATHODE MERCURY–ARC POWER CONVERTERS. *ASA C34.1–1958*, American Standards Association, Inc., New York, N. Y., 1958.

13. HARMONIC ANALYSIS BY DIRECT AREA MEASUREMENTS, **I. K. Dortort.** *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 75, Mar. 1956, pp. 16–19.

14. ANALYSIS OF RECTIFIER CIRCUITS, **E. F. Christensen, C. H. Willis, C. C. Herskind.** *Ibid.*, vol. 63, 1944, pp. 1048–58.

15. THE CALCULATION OF RECTIFIER AND INVERTER PERFORMANCE CHARACTERISTICS, **J. C. Read.** *Journal*, Institution of Electrical Engineers, London, England, vol. 92, pt. II, no. 29, Oct. 1945.

Discussion

I. K. Dortort (I-T-E Circuit Breaker Company, Philadelphia, Pa.): The author has applied thought and effort in attacking an old problem from a new angle. Any novel method has academic value, but the choice of aiming point is seriously questioned.

The stated end result of the paper is an rms value of the ripple voltage at the output terminals of a rectifier, from which it is hoped to evaluate:

1. Ripple noise in the d-c load.
2. Telephone interference from the d-c lines.
3. Inductive heating in the d-c system.
4. A-c losses in the d-c load.

Items 2, 3, and 4 are definitely related to the ripple current rather than the ripple voltage. Item 1 may be influenced by either one or both. Another important effect, not mentioned by the author, is the peak-to-peak ripple component of magnetic flux as in particle accelerators, focusing coil, bubble chambers, etc. Here again, the principal concern is generally the amplitude of ripple current.

Except in the case of capacitive loads or purely resistive systems (including the rectifier unit), the principal ripple frequency current and its harmonics are increasingly attenuated as the frequency increases, by any inductance present anywhere in the d-c path. Moreover, as the harmonic components of the ripple voltage generally decrease with increasing order of the harmonics, and the impedance of the circuit increases in the same direction, the harmonic components of the current rapidly become negligible with increasing order. Obviously, then, the total rms value of the ripple current is not proportional to the ripple voltage.

Curves are available from several sources, showing the rms values of the more important harmonics of the d-c ripple voltage as a function of load and phase control. Because of the decreasing value of the higher harmonics it is not much of a chore to obtain a fairly accurate value of the rms ripple voltage, and a much more accurate value of the ripple current.

If a single figure of merit is considered desirable, it is suggested that the half-cycle average or the time integral of the ripple voltage over one half-cycle of the principal ripple frequency are more easily obtained than the rms values, and far more useful. Determination by integration or by direct area measurements is easily accomplished.

The use of the volt-second value of the ripple voltage leads immediately to useful results. In an inductive circuit in which inductance-resistance is relatively large

$$(i_{\max}-i_{\min})=\frac{1}{L}\int edt$$

This can be converted to total rms with good accuracy by the factor $1/(2\sqrt{2})$.

For the general case we can write

$$\int edt=\sum_{n=1}^{\infty} e_{n_q}dt=L(i_{\max}-i_{\min})+$$

$$\sum_{n=1}^{\infty}\frac{1}{\sqrt{1+\left(\frac{nq\omega L}{R}\right)^2}}\int e_{n_q}dt$$

For a moderately inductive circuit this expression reduces with fair accuracy to

$$(i_{\max}-i_{\min})=\frac{1}{L}\left(1-\frac{1}{\sqrt{1+\left(\frac{q\omega L}{R}\right)^2}}\right)\int edt$$

since the n terms of the summation decrease rapidly for values of n greater than 1.

E. J. Diebold: Mr. Dortort's contribution is appreciated; an authority in the field, he obviously does not need a novel or simplified approach to this old problem.

Current industrial practice, and proposed standards, refer to the per-cent ripple or the per-unit ripple as an effective value. It may be desirable to follow the proposed method using the "half-cycle average." This, however, is far beyond the scope of the paper.

Manufacturers and users of rectifier equipment containing new circuits for other nonstandard approaches are faced with the necessity of rapidly finding a generally accepted value for the magnitude of the ripple, preferably without much numerical work. It has been our experience that the method shown in this paper yields acceptable results with acceptable accuracy within a short time. The results, therefore, were published to introduce a broader public to this simplified method.

Minimization of Multiple-Output Switching Circuits

R. B. POLANSKY
NONMEMBER AIEE

THERE is a variety of systematic procedures for simplifying single-output switching functions. This paper proposes an extension of a certain class of these procedures. This extension makes it possible to simplify multiple-output switching circuits by using tags to identify the terms belonging to different output functions.

In particular, the Quine-McCluskey[1] procedure, as described by Caldwell,[2] can be extended by this method. This extended procedure can be used to determine the minimal multiple-output switching circuit. For the example to be given, the minimal switching circuit is defined as the 2-level AND-OR diode-resistor circuit which uses the fewest diodes.

Other single-output simplification procedures, such as the Roth procedure,[3] can also be extended in a similar manner. The object in extending these other procedures is to obtain different types of minimal circuits. For example, the Roth procedure can be used to obtain circuits which have more than two levels.

In this paper, the symbol + signifies disjunction, the "logical OR" operation. Juxtapositions, such as $x_1x_2x_3$, are usually called "conjuncts," or "product terms," and the binary literals x_i^*, $1\leq i\leq p$, are used here to represent the p inputs to the switching circuit. In this paper x_i^* has two possible values, x_i, or $\bar{x}_i$, the complement of x_i. These symbols are occasionally replaced by binary digits 0 and 1. In this paper, 0 sometimes replaces $\bar{x}_i$ and 1 sometimes replaces x_i.

The Multiple-Output Model

A multiple-output switching circuit S with q outputs may be described by an "ensemble" $E[F_i]$ of q switching functions $F_1, F_2, \ldots, F_q$, one for each output of S. It is assumed, in the method to be described, that each F_i is given as a canonical sum of products of the p input literals $x_1^*, x_2^*, \ldots, x_p^*$. A typical 3-input 3-output function is given in canonical form in Table I.

Definition and Use of Tags

It is possible for several output functions $F_a, F_b, \ldots$, etc., of $E[F_i]$ to have some identical canonical product terms. These

Table I. A Multiple-Output Switching Function in Canonical Sum of Products Form

$$F_1=\bar{x}_1x_2x_3+x_1x_2x_3+\bar{x}_1x_2\bar{x}_3+x_1\bar{x}_2x_3+\bar{x}_1\bar{x}_2x_3=\Sigma(1,2,3,5,7)$$

$$F_2=x_1x_2x_3+x_1x_2\bar{x}_3+\bar{x}_1x_2\bar{x}_3+x_1\bar{x}_2\bar{x}_3+\bar{x}_1\bar{x}_2\bar{x}_3=\Sigma(0,2,4,6,7)$$

$$F_3=\bar{x}_1x_2x_3+\bar{x}_1x_2\bar{x}_3+x_1x_2x_3+x_1x_2\bar{x}_3+\bar{x}_1\bar{x}_2\bar{x}_3+x_1\bar{x}_2\bar{x}_3=\Sigma(0,2,3,4,6,7)$$

Paper 60-1223, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960. Manuscript submitted June 10, 1960, made available for printing October 11, 1960.

R. B. POLANSKY is with the Electronic Systems Laboratory of the Massachusetts Institute of Technology, Cambridge, Mass.

The research reported in this article was made possible through the support extended the Massachusetts Institute of Technology, Electronic Systems Laboratory, by the United States Air Force (Weapons Guidance Laboratory, Wright Air Development Division), under Contract No. *AF-33(616)-5477*. It is published for technical information only and does not represent recommendations or conclusions of the sponsoring agency.

terms can be realized by one circuit which is "OR-ed" to each of the respective circuit output terminals. This subcircuit sharing by several outputs may result in a simpler, more desirable circuit than would otherwise be obtained.

Tags are used to denote this interfunctional sharing. The tags, which are used for a q-output switching ensemble, consist of q binary digits $T_1^*, \ldots, T_q^*$. The convention is that T_i^* has the value 0 (i.e., $\bar{T}_i$) if the tagged term is a member of F_i, or T_i^* has the value 1 (i.e., T_i) if the tagged term is not a member of F_i. Thus, any single canonical product term would have a tag with only a zero digit, while a term combining canonical product terms from several F_i would have a zero tag digit for each of them. Since a given conjunct can be a member of any combination of F_i, and there are 2^q of these combinations, it follows that the q binary digits of the tags are sufficient to designate any possible combination of F_i of which a conjunct may be a member. Also, this type of tag can conveniently be manipulated by any digital computer which uses binary logic.

Some examples of these tags illustrate the previous description. The fourth product term of F_1 in Table I is tagged as follows: $\bar{T}_1T_2T_3x_1\bar{x}_2x_3$. Such a tagged canonical product term is called a "p-term" of $E[F_i]$. Any term which denotes a combination of several p-terms is called an "s-term." For example, the two p-terms $\bar{T}_1T_2T_3\bar{x}_1x_2x_3$ and $T_1T_2\bar{T}_3\bar{x}_1x_2x_3$ can be combined. The combination s-term is written $\bar{T}_1T_2\bar{T}_3\bar{x}_1x_2x_3$, which means that the circuit for $\bar{x}_1x_2x_3$ is shared by outputs 1 and 3. The nontag digits (i.e., x_1^*, $x_2^*, \ldots, x_p^*$) of any p-term or s-term are referred to as the "kernel."

There is a second type of s-term which also denotes a group of p-terms. In these the kernel is modified to show that the group of p-terms belongs to the same output function. For example, $\bar{T}_1T_2T_3x_1\text{-}x_3$ is the combination of the two p-terms $\bar{T}_1T_2T_3x_1x_2x_3$, and $\bar{T}_1T_2T_3x_1\bar{x}_2x_3$. Similarly $\bar{T}_1T_2\bar{T}_3-x_2x_3$ is the s-term which stands for the combination of the two s-terms, $\bar{T}_1T_2T_3-x_2x_3$ and $T_1T_2\bar{T}_3-x_2x_3$, or also of $\bar{T}_1T_2T_3x_1x_2x_3$ and $\bar{T}_1T_2T_3\bar{x}_1x_2x_3$. The blank signifies that the x_i^* for that position may have either value x_i or $\bar{x}_i$.

Covers

A necessary condition for the final disjunction to be a solution is that it accurately represents the given functions $E[F_i]$. This condition is fulfilled if the disjunction is a cover as defined in the following.

Let TK stand for any p-term or s-term, where T is its tag and K is its kernel. Further, T_{ki}^* is digit T_k^* of the term $(TK)_i$ and similarly K_{hi}^* is digit x_h^* of the term $(TK)_i$. The term $(TK)_j$ is said to ***imply*** another term $(TK)_i$, denoted by $(TK)_j \rightarrow (TK)_i$, if and only if the following two conditions are both fulfilled:

1. $T_{kj}^*=0$, if $T_{ki}^*=0$ for all k, where $1 \leq k \leq q$
2. If K_{hj}^* is not a blank $(-)$ then $K^*_{hj}=K_{hi}^*$ for all h, where $1 \leq h \leq p$.

There are no conditions for $T_{ki}^*=1$ or $K_{hj}^*=(-)$. Thus $\bar{T}_1T_2T_3x_1-x_3 \rightarrow \bar{T}_1T_2T_3x_1\bar{x}_2x_3$.

A disjunction of N-tagged conjuncts is symbolized by

N
$D(TK)_j$
$j=1$

Such a disjunction is said to be a "cover" V of $E[F_i]$ if and only if:

1. For every p-term $(TK)_p$ of $E[F_i]$ there is at least one $j, 1 \leq j \leq N$, such that $(TK)_j \rightarrow (TK)_p$.
2. For every $(TK)_j$ of the disjunction, and every $(TK)_h$ such that $(TK)_j \rightarrow (TK)_h$ and $(TK)_h$ has the form of a p-term (i.e., $(TK)_h$ has a single zero tag digit and no blank kernel digits), then $(TK)_h$ is indeed a p-term of $E[F_i]$. This condition assures that the cover V includes only p-terms.

For example, the ensemble shown in Table II along with its p-terms, is covered by the disjunction

$$\bar{T}_1\bar{T}_2x_1x_2x_3+\bar{T}_1T_2x_1x_2-+T_1\bar{T}_2x_1-x_3$$

for

$\bar{T}_1\bar{T}_2x_1x_2x_3 \longrightarrow \bar{T}_1T_2x_1x_2x_3$	p-term 1
$\bar{T}_1\bar{T}_2x_1x_2x_3 \longrightarrow T_1\bar{T}_2x_1x_2x_3$	p-term 3
$\bar{T}_1T_2x_1x_2- \longrightarrow \bar{T}_1T_2x_1x_2x_3$	p-term 1
$\bar{T}_1T_2x_1x_2- \longrightarrow \bar{T}_1T_2x_1x_2\bar{x}_3$	p-term 2
$T_1\bar{T}_2x_1-x_3 \longrightarrow T_1\bar{T}_2x_1x_2x_3$	p-term 3
$T_1\bar{T}_2x_1-x_3 \longrightarrow T_1\bar{T}_2x_1\bar{x}_2x_3$	p-term 4

Thus, condition 1 is fulfilled because every p-term is implied, and condition 2 is fulfilled because only p-terms are implied.

Cost Functions

The object of any simplification procedure is to obtain the circuit which performs the desired logical operation at the smallest cost. Thus, to be able to find a minimal cost switching circuit, it is necessary to have a cost criterion.

Let $C(V)$ be the cost function which equals some cost associated with constructing switching circuit S as described by the cover V. For the example to be given, the cost function equals the number of diodes needed to build a 2-level diode-resistor AND–OR switching circuit. To derive this cost function, consider the number of diodes needed to include the circuit for any term, TK, of the cover in the final circuit. Nelson[4] has described these AND–OR diode-resistor circuits. One diode is needed for each input to the first-level AND gate. There are L of these inputs, one for each nonblank kernel literal. Also, one diode is needed for each second-level OR gate in which the conjunct is used. There are Z of these, one for each zero tag digit. But, if one conjunct completely covers an output function, then no output OR gate is needed for that function, and the cost Z may be reduced by β, the number of functions which TK completely covers. Finally, if there is only one input to the first level, no AND gate is needed. Thus, L may be reduced from 1 to 0. This is done by subtracting γ from the cost, where $\gamma=1$ if $L=1$, otherwise $\gamma=0$. The cost function for counting diodes is

$$C(TK)=L+Z-\beta-\gamma\left\{\begin{array}{l}\gamma=1 \text{ if } L=1\\ \gamma=0 \text{ if } L\neq 1\end{array}\right.$$

Table II. An Ensemble and p-Terms

Ensemble	p-Terms	
$F_1=x_1x_2x_3+x_1x_2\bar{x}_3$	1	$\bar{T}_1T_2x_1x_2x_3$
$F_2=x_1x_2x_3+x_1\bar{x}_2x_3$	2	$\bar{T}_1T_2x_1x_2\bar{x}_3$
	3	$T_1\bar{T}_2x_1x_2x_3$
	4	$T_1\bar{T}_2x_1\bar{x}_2x_3$

Table III. Multiple-Output Switching Function Used in the Example

$F_1=x_1x_2x_3x_4+x_1x_2x_3\bar{x}_4=\Sigma(14,\ 15)$
$F_2=\bar{x}_1x_2\bar{x}_3\bar{x}_4+x_1x_2\bar{x}_3\bar{x}_4+x_1x_2\bar{x}_3x_4=\Sigma(4,\ 12,\ 13)$
$F_3=x_1x_2x_3x_4+x_1\bar{x}_2x_3x_4+\bar{x}_1\bar{x}_2x_3x_4=\Sigma(3,\ 11,\ 15)$
$F_4=\bar{x}_1x_2\bar{x}_3\bar{x}_4+\bar{x}_1\bar{x}_2x_3\bar{x}_4+\bar{x}_1x_2x_3\bar{x}_4=\Sigma(2,\ 4,\ 6)$
$F_5=x_1x_2x_3x_4+x_1x_2x_3\bar{x}_4+\bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4=\Sigma(0,\ 14,\ 15)$
$F_6=x_1x_2x_3x\ +x_1\bar{x}_2x_3x\ +\bar{x}_1\bar{x}_2x_3x\ +\bar{x}_1\bar{x}_2\bar{x}_3\bar{x}$
$=\Sigma(0,\ 3,\ 11,\ 15)$

Table IV. The List of p-Terms

Number	Tag	Kernel
1	011111	1111
2	011111	1110
3	101111	0100
4	101111	1100
5	101111	1101
6	110111	1111
7	110111	1011
8	110111	0011
9	111011	0100
10	111011	0010
11	111011	0110
12	111101	1111
13	111101	1110
14	111101	0000
15	111110	1111
16	111110	1011
17	111110	0011
18	111110	0000

Table V. List L_0

Tag	Kernel	
111100	0000	*a*
111011	0010	√
101011	0100	*b*
110110	0011	√
111011	0110	√
101111	1100	√
110110	1011	√
101111	1101	√
011101	1110	√
010100	1111	*c*

Table VI. List L_1

Tag	Kernel	
111011	0–10	*d*
111011	01–0	*e*
101111	–100	*f*
110110	–011	*g*
101111	110–	*h*
110110	1–11	*i*
011101	111–	

Table VII. Prime Implicants Derived From $\bar{T}_1\bar{T}_2\bar{T}_3\bar{T}_4x_1\text{-}x_3$

Tag	Kernel
0101	1–1
0110	1–1
1100	1–1
0111	1–1
1101	1–1
1110	1–1

and finally

$$C(V) = \sum_{(TK)_j \epsilon V} C(TK)_j$$

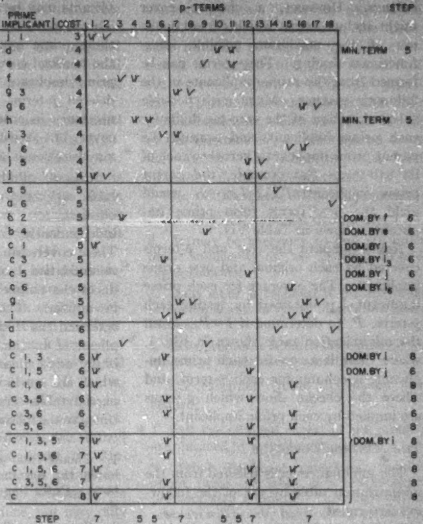
A cover V' is said to be a "minimal cover" of $E[F_i]$ if and only if $C(V') \leq C(V)$ where V is any other cover of $E[F_i]$. The minimal switching circuit S' is described by V'. Hence, the procedure, described next, is one which will obtain V' for any given $E[F_i]$.

The Minimization Procedure

The expanded procedure can briefly be described as a sequence of three operations. First, the functions are transformed by the use of tags so that they can be manipulated simultaneously in the same manner as a single-output function.

Second, all possible combinations of p-terms are made within the different output functions and also by sharing between different output functions. Finally, the disjunction of some of these combinations is chosen which covers $E[F_i]$ at the lowest cost. This disjunction describes the minimal switching circuit. These operations are described in greater detail, and are illustrated by the following example.

Fig. 1. Minimization table



PRIME IMPLICANT	COST	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		STEP
j 1	3	✓	✓																		
d	4										✓	✓								MIN. TERM	5
e	4									✓		✓									
f	4			✓	✓																
g 3	4							✓	✓												
g 6	4																✓	✓			
h	4				✓	✓														MIN. TERM	5
i 3	4						✓	✓													
i 6	4															✓	✓				
j 5	4												✓	✓							
j	4	✓	✓										✓	✓							
a 5	5														✓						
a 6	5																		✓		
b 2	5			✓																DOM. BY f	6
b 4	5									✓										DOM. BY e	6
c 1	5	✓																		DOM. BY j	6
c 3	5						✓													DOM. BY i_3	6
c 5	5												✓							DOM. BY j	6
c 6	5															✓				DOM. BY i_6	6
g	5							✓	✓								✓	✓			
i	5						✓	✓								✓	✓				
a	6														✓				✓		
b	6			✓						✓											
c 1, 3	6	✓					✓													DOM. BY i	8
c 1, 5	6	✓											✓							DOM. BY j	6
c 1, 6	6	✓														✓				DOM. BY i	8
c 3, 5	6						✓						✓								8
c 3, 6	6						✓									✓					6
c 5, 6	6												✓			✓					8
c 1, 3, 5	7	✓					✓						✓								8
c 1, 3, 6	7	✓					✓									✓					8
c 1, 5, 6	7	✓											✓			✓					8
c 3, 5, 6	7						✓						✓			✓					8
c	8	✓					✓						✓			✓					8
STEP		7			5	5		7			5	5	7				7				

p-TERMS

NOTE: A REMOVED ROW OR COLUMN IS DESIGNATED BY CROSSING THE CHECK(✓) IN IT.

An Example: The Extended Quine-McCluskey Procedure

In order to illustrate the Extended Quine-McCluskey Procedure, find the minimal cover, as defined previously, for the 4-input 6-output switching function given in Table III.

Step 1: Tag the canonical product terms, and number them sequentially for future reference. The p-terms, formed in this manner, are shown in Table IV, where the variables are replaced by their binary equivalents, as mentioned in the introduction.

Step 2: Form the list L_0 by combining each group of p-terms with identical kernels into a new s-term with the same kernel. The tag for this new term is formed by "logically multiplying" corresponding digits in the tags of the combining p-terms. Any p-term with a kernel that occurs only once is entered unchanged into L_0. Thus, the disjunction of terms in L_0 is a cover for $E[F_i]$ and shows the maximum possible interfunction sharing. List L_0 is shown in Table V. The tagged conjuncts in L_0 are arranged in increasing order of the number of ones in their kernels, as is done in the Quine-McCluskey procedure.

Step 3: This step corresponds directly to the Quine-McCluskey procedure. Lists $L_1, \ldots, L_n$ are formed using the rule $x_jx_k + \bar{x}_jx_k = -x_k$ to combine kernels. The rule for combining tags is $T_{1i}{}^* + T_{2i}{}^* = T_i{}^*$ $i = 1, 2, \ldots, q$, where $T_{1i}{}^*$ and $T_{2i}{}^*$ are the ith digits of the tags of the combining terms $(TK)_1$ and $(TK)_2$, and $T_i{}^*$ is the i^{th} digit of the new tag. (In other words, the new tag is formed by logically adding corresponding digits in the combining tags.) List L_1 is shown in Table VI. However, no further combining is possible in this particular example.

Since the minimal cover is a disjunction of only some of the s-term and p-terms given in lists $L_0, \ldots, L_n$, it may be possible to eliminate several of these terms as possible mcv terms (terms appearing in the minimal cover which are referred to as *MIN TERMS* in the figures) just as soon as they are formed. For if $(TK)_k \rightarrow (TK)_m$ and $C(TK)_k < C(TK)_m$, then $(TK)_m$ cannot possibly be a mcv term. Hence, any combining term which is implied by a cheaper combined term may be marked with a √, signifying that it is not a mcv term. Any term with all ones in its tag may be removed, since it is not used in any F_i.

Step 4: The minimal cover is to be selected from the remaining unmarked terms, called "prime implicants." These prime implicants are designed with lower case letters for future reference.

The tags show the entire set of output functions to which each prime implicant belongs. However, a cheaper cover might include terms which do not make use of the maximum possible inter-functional sharing. These terms can be formed from the prime implicants in the following manner. Make each possible subcombination of the zero tag digits for each prime implicant, and append the parent prime implicant's kernel to each of its sub-tags. For example, the parent prime implicant $T_1T_2T_3T_4x_1-x_3$ would yield the new terms (also prime implicants) shown in Table VII.

Next, compute the cost and *p*-terms covered by each original and new prime implicant. The coverage for each prime implicant *I* is obtained by testing each *p*-term, *P*, to determine if $I \rightarrow P$. Form the minimization table shown in Fig. 1, where there is a row for each prime implicant, a column for each *p*-term, and where the checks show which *p*-terms are implied by each prime implicant.

THE MINIMAL COVER

The minimal cover is selected from the minimization table by use of the following five rules:

Rule 1: If there is only one check in any column, the prime implicant whose row the check appears in is needed to cover the function, and therefore, is a mcv term. The columns corresponding to all of the *p*-terms which that mcv term implies may be removed from the table. The mcv term is placed in the mcv term list and its row may also be removed from the table.

Rule 2: A prime implicant *P* is said to dominate another prime implicant *Q* if *P* covers at least the same remaining *p*-terms as *Q*, and if $C(P)<C(Q)$. In this situation the row for *Q* may be removed from the table.

Rule 3: Whenever a *p*-term P_p is implied by only a small number of prime implicants, then any other *p*-terms which are implied by all of these prime implicants may be removed from the table. This is possible because at least one of these prime implicants must appear in the minimal cover to imply P_p, and this prime implicant is certain to imply the deleted *p*-terms also. This rule is not necessary in order to find the minimum cover, but it will frequently reduce the minimization labor.

PRIME IMPLICANT	COST	2	3	6	8	9	13	14	15	17	18
j 1	3	✓									
e	4					✓					
f	4		✓								
g 3	4				✓						
g 6	4									✓	
i 3	4			✓							
i 6	4								✓		
j 5	4						✓				
j	4	✓					✓				
a 5	5							✓			
a 6	5										✓
g	5				✓					✓	
i	5			✓					✓		
a	6							✓			✓
b	6		✓			✓					

Fig. 2. Reduced minimization table

Rule 4: Prebranching minimization table separation: It may be possible to obtain covers for some groups of *p*-terms independently of the other *p*-terms. These covers can be found from a partial minimization table, the use of which reduces the number of times Rule 5 must be applied. The partial table is characterized as follows: Any prime implicant *I* in the partial table does not imply any remaining uncovered *p*-terms which are not in this partial table. The partial tables are determined, one at a time, by the following procedure.

1. Choose all prime implicants which imply the lowest numbered remaining *p*-term. Enter these prime implicants in list *A*.

2. Enter in list *B* all the *p*-terms which are implied by any member of list *A*. Check (√) the lowest numbered *p*-term in list *B*. The check mark indicates that all prime implicants which imply the checked *p*-term have been entered in list *A*.

3. Choose the lowest numbered unchecked *p*-term in list *B*, check it, and place in list *A* any prime implicants which imply this *p*-term and are not already in list *A*. Any *p*-terms which these new prime implicants also imply and which are not already entered in list *B* should be entered there now.

Step 3 should be repeated until all *p*-terms in list *B* have been checked. Lists *A* and *B* now define an independent portion of the minimization table to which Rule 5 will be applied. Call this portion the Current Minimization Table (CMT). Of course, there may not be any independent groups of *p*-terms and prime implicants. In this case, list *B* will contain all of the *p*-terms and list *A* will contain all of the prime implicants.

Rule 5: Branching: Choose the prime implicant which implies the largest number of *p*-terms in list *B* of Rule 4, and assume it is a mcv term. The *p*-terms implied by it may be removed from CMT and the mcv term selection continues as before, starting with Rule 1. Next, that prime implicant is assumed not to be a mcv term. Delete it from list *A* of Rule 4, and continue minimizing, starting with Rule 1. This produces two possible covers. The costs are compared, and the less expensive is the minimum cover. Note that several successive branches may be necessary, in which case all possible combinations of "branched-on" prime implicants must be compared. The least expensive combination is part of the minimum cover. Because Rule 5 requires several alternate covers to be found, it should not be applied until Rules 1, 2, and 3 no longer help simplify the minimization table.

PRIME IMPLICANT	COST	2	13
j 1	3	✓	
j 5	4		✓
j	4	✓	✓

(A)

	EXAMPLE NO. OF DIODES
INDEPENDENT MINIMIZATION	37
MULLER SOLUTION 1	35
MULLER SOLUTION 2	37
DIODE MINIMIZATION $C=L+Z-\beta-\gamma$	㉞

(B)

Fig. 3

A—Current minimization table for step 9
B—Cost comparison for different simplification procedures

If all *p*-terms of $E[F_i]$ are not yet implied by mcv terms, then Rule 4 should be applied to the remaining minimization table, which has a column for each uncovered *p*-term.

A possible ordering for applying these rules is given in the Appendix as a set of flow charts (Figs. 4–6).

CONTINUATION OF EXAMPLE

The application of these five rules of minimal cover will now be illustrated by means of the example previously started.

Step 5: Select the mcv terms according to Rule 1. These terms are *d* and *h*.

Step 6: Remove the dominated prime implicants according to Rule 2. These are *b*2, *b*4, *c*1, *c*3, *c*5, *c*6, *c*1, 5 and *c*3, 6.

Step 7: Since Rule 1 does not produce any more mcv terms, apply Rule 3. Start with column 2. Column 1 may now be removed since *j*1 and *j* both imply *p*-term 1. Similarly, columns 7, 12, and 16 may also be removed.

Step 8: Apply Rule 2. Prime implicants *c*1,3; *c*,6; *c*3,5; *c*5,6; *c*1,3,5; *c*1,3,6; *c*1,5,6; *c*3,5,6 and *c* may be removed. The remaining minimization table is shown in Fig. 2.

Step 9: Rules 1, 2, and 3 do not produce further simplication of the minimization table. Rule 4 is applied, as shown in Table VIII.

Table VIII. Lists Used in Step 9

List A Prime Implicants	List p-Terms
*j*1 (*a*)	2 (*b*)
j (*a*)	13 (*c*)
*j*5 (*c*)	

Table IX. Minimal Cover for the Example

$F_1 = x_1x_2x_3-$
$F_2 = \bar{x}_1x_2\bar{x}_3\bar{x}_4 + x_1x_2\bar{x}_3-$
$F_3 = -\bar{x}_2x_3x_4 + x_1-x_3x_4$
$F_4 = \bar{x}_1x_2\bar{x}_3\bar{x}_4 + \bar{x}_1-x_3\bar{x}_4$
$F_5 = \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4 + x_1x_2x_3-$
$F_6 = \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4 + -\bar{x}_2x_3x_4 + x_1-x_3x_4$

Step 10: Apply Rule 5 to the portion of the CMT determined in step 9 and shown in Fig. 3(A). Prime implicant *j* covers the most *p*-terms, so *j* is selected. This is a cover of cost 4. The branch without *j* requires both *j*1 and *j*5 by Rule 1. This cover has a cost of 7. Hence, *j* is a mcv term.

Step 11: Apply Rule 4 to the remaining portion of the minimization table (Fig. 2 and columns 2 and 13 removed). This gives:

List A	List B
f	3√
b	9√
e	

and by Rule 5, *b* is found to be a mcv term.

Step 12: The procedure of steps 9, 10, and 11 is continued and *i*, *g*, and *a* are determined as the remaining mcv terms.

The minimal 2-level diode-resistor switching circuit has a cost of 34 diodes and is described by the disjunction of prime implicants $a+b+d+g+h+i+j$. Table IX shows this unique solution in sum-of-products form.

Other Methods

There are other methods for simplifying multiple-output switching circuits. Although the two to be discussed next have certain advantages, they do not always provide the minimal circuit, which is obtained by the tagging procedure.

Method 1. Simplify each F_i independently, using an existing single-output simplification procedure. This method is easier and less time consuming to perform than the simultaneous simplication procedures. However, it may not obtain the minimal cover, since no allowance is made for interfunctional sharing.

Method 2. The Muller Method.[5] This method makes allowance for interfunctional sharing, but the implicitly stated cost function does not guarantee that the minimal cover will be obtained. Also, a large number of "don't care" conditions are used which greatly increase the labor.

Method 3. The extended Quine-McCluskey procedure. This method does not require significantly more labor for forming the prime implicants than Method 1, and the use of a definite cost function guarantees that the minimal cover will be obtained, but the minimal cover selection is considerably more work than in Method 1.

A comparison of the covers obtained by these methods is shown in Fig. 3(B). The Muller procedure produced two different covers. The circled number is the minimum cost. Only the diode minimization procedure obtained the best 2-level diode-resistor circuit for the example given in Table III.

Method 4. A method has been described by R. McNaughton and B. Mitchell[6] for obtaining the multiple-output rectifier net with the fewest rectifiers. This method will obtain at least one of the minimal covers found by the extended Quine-McCluskey procedure. It is important to point out the difference between what was termed a prime implicant in this paper, and what was termed an *r*th order prime implicant by McNaughton and Mitchell. A prime implicant need not cover any *p*-terms of the indicated output functions, but may only cover "don't care" terms in one or more of these output functions. This is a violation of conditions 3 and 3′ in the definition of an *r*th order prime implicant. Also, a prime implicant need not fulfill conditions 4, 4′, 5, or 5′ of an *r*th order prime implicant. This means that there are more terms available from which to minimal covers. However, any terms which will not be needed in the alternate minimal covers are quickly eliminated in minimization steps 5 and 6 of the extended Quine-McCluskey procedure. It should be remembered that only prime implicants which are used in one of the minimal covers are referred to as mcv terms. The following is a brief summary of these two methods.

The McNaughton-Mitchell procedure:

1. At least one minimal cover is obtained.
2. It may be necessary to find *r*th order prime implicants for 2^q-1, single-output-type problems.
3. Less labor is required if the *r*th order prime implicants have few literals and many blanks.
4. The procedure is much more difficult for arbitrary "don't care" conditions.

The extended Quine-McCluskey procedure:

1. All the minimal covers of the problem are obtained by a systematic procedure.
2. Prime implicant generation requires about twice the labor of a single-output problem with *p* inputs. This single-output problem will usually have more *p*-terms than any one of the 2^q-1 problems performed in the McNaughton-Mitchell procedure.
3. Less labor is needed if the prime implicants have many literals and few blanks.
4. The method allows for arbitrary "don't care" conditions.

Note that the Quine-McCluskey pro-

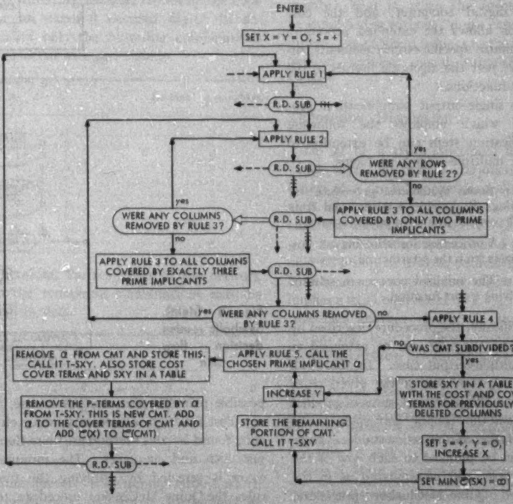

ENTER
SET X = Y = O, S = +
APPLY RULE 1
R.D. SUB
APPLY RULE 2
R.D. SUB
yes
WERE ANY ROWS REMOVED BY RULE 2?
no
APPLY RULE 3 TO ALL COLUMNS COVERED BY ONLY TWO PRIME IMPLICANTS
yes
WERE ANY COLUMNS REMOVED BY RULE 3?
R.D. SUB
no
APPLY RULE 3 TO ALL COLUMNS COVERED BY EXACTLY THREE PRIME IMPLICANTS
R.D. SUB
yes
WERE ANY COLUMNS REMOVED BY RULE 3?
no
APPLY RULE 4
no
WAS CMT SUBDIVIDED?
yes
REMOVE α FROM CMT AND STORE THIS. CALL IT T-SXY. ALSO STORE COST COVER TERMS AND SXY IN A TABLE
APPLY RULE 5. CALL THE CHOSEN PRIME IMPLICANT α
INCREASE Y
STORE SXY IN A TABLE WITH THE COST AND COVER TERMS FOR PREVIOUSLY DELETED COLUMNS
REMOVE THE P-TERMS COVERED BY α FROM T-SXY. THIS IS NEW CMT. ADD α TO THE COVER TERMS OF CMT AND ADD Ɛ(X) TO Ɛ(CMT)
STORE THE REMAINING PORTION OF CMT. CALL IT T-SXY
SET S = +, Y = O, INCREASE X
R.D. SUB
SET MIN Ɛ(SX) = ∞

Fig. 4. Flow chart showing rule application ordering

cedure was extended as an example and is only one of several procedures which can be extended to the multiple-output case by means of the tagging procedure given in this paper.

"Don't-Care" Conditions

In the extended Quine-McCluskey procedure "don't-care" conditions are tagged and manipulated as if they are *p*-terms, but no columns are included for them in the minimization table. The definition of "implies" and also of "cover" must be modified as follows:

$(TK)_j$ implies $(TK)_i$ where there are don't cares (symbolized by $(TK)_j \xrightarrow{dc} (TK)_i$) if for every *p*-term $(TK)_p$ such that $(TK)_i \rightarrow (TK)_p$ then $(TK)_j \rightarrow (TK)_p$ also.

Condition 2 of the definition of cover must be modified to read:

For every $(TK)_j$ of the disjunction, and every $(TK)_a$ such that $(TK)_j \xrightarrow{dc} (TK)_a$ and $(TK)_a$ has the form of a *p*-term (i.e., $(TK)_a$ has a single zero tag digit and no blank kernel digits), then $(TK)_a$ is either a *p*-term or a don't-care term of $E[F_i]$.

Conclusions

The tagging procedure allows the simultaneous simplification of multiple-output switching circuits without greatly increasing the dimension of the problem. The tag manipulations are simple, so that they can be conveniently performed by a digital computer, and the cost function allows the extended procedure to minimize specific circuit elements, instead of just the algebraic literals in the output functions.

Any single-output simplification procedure which employs the following simplification steps can be extended to handle multiple-output problems.

Step 1: Some systematic procedure for generating all possible *p*-terms, and from them all possible simplified terms (*s*-terms).

Step 2: A procedure for selecting all possible covers from the *p*-terms and *s*-terms.

Step 3: The minimal covers are selected by applying a cost function.

Specifically, the procedure described by Roth and Wagner[3] can be extended to the multiple-output case. In the Roth procedure, the rows of the given truth table are the canonical sums-of-product terms, and the "scopes" correspond to the rows of the minimization table. Thus, tags can be prefixed to each row of the truth table, and manipulated as in the extended Quine-McCluskey procedure, when the truth table is modified to form possible prime implicants. Of course, a new cost function is necessary, since circuits found by this method use more than just AND and OR gates. The minimal cover is selected by applying the five rules, the Quine-McCluskey procedure, to the "scopes."

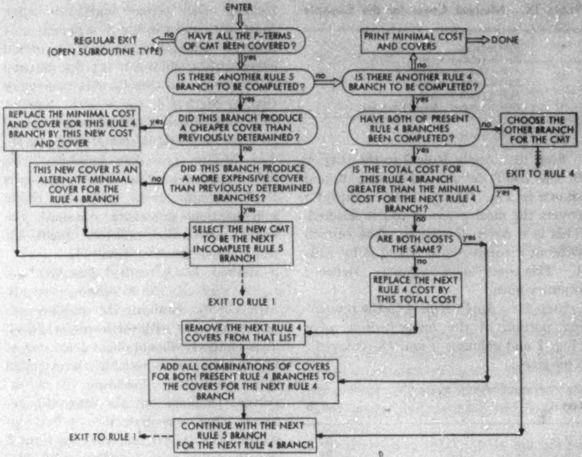

ENTER
HAVE ALL THE P-TERMS OF CMT BEEN COVERED?
no
REGULAR EXIT (OPEN SUBROUTINE TYPE)
yes
IS THERE ANOTHER RULE 5 BRANCH TO BE COMPLETED?
no
IS THERE ANOTHER RULE 4 BRANCH TO BE COMPLETED?
no
PRINT MINIMAL COST AND COVERS
DONE
yes
DID THIS BRANCH PRODUCE A CHEAPER COVER THAN PREVIOUSLY DETERMINED?
yes
REPLACE THE MINIMAL COST AND COVER FOR THIS RULE 4 BRANCH BY THIS NEW COST AND COVER
no
DID THIS BRANCH PRODUCE A MORE EXPENSIVE COVER THAN PREVIOUSLY DETERMINED BRANCHES?
no
THIS NEW COVER IS AN ALTERNATE MINIMAL COVER FOR THE RULE 4 BRANCH
yes
SELECT THE NEW CMT TO BE THE NEXT INCOMPLETE RULE 5 BRANCH
EXIT TO RULE 1
yes
HAVE BOTH OF PRESENT RULE 4 BRANCHES BEEN COMPLETED?
no
CHOOSE THE OTHER BRANCH FOR THE CMT
EXIT TO RULE 4
yes
IS THE TOTAL COST FOR THIS RULE 4 BRANCH GREATER THAN THE MINIMAL COST FOR THE NEXT RULE 4 BRANCH?
yes
no
ARE BOTH COSTS THE SAME?
yes
no
REPLACE THE NEXT RULE 4 COST BY THIS TOTAL COST
REMOVE THE NEXT RULE 4 COVERS FROM THAT LIST
ADD ALL COMBINATIONS OF COVERS FOR BOTH PRESENT RULE 4 BRANCHES TO THE COVERS FOR THE NEXT RULE 4 BRANCH
CONTINUE WITH THE NEXT RULE 5 BRANCH FOR THE NEXT RULE 4 BRANCH
EXIT TO RULE 1

Fig. 5 (above). Verbal routing decision flow chart

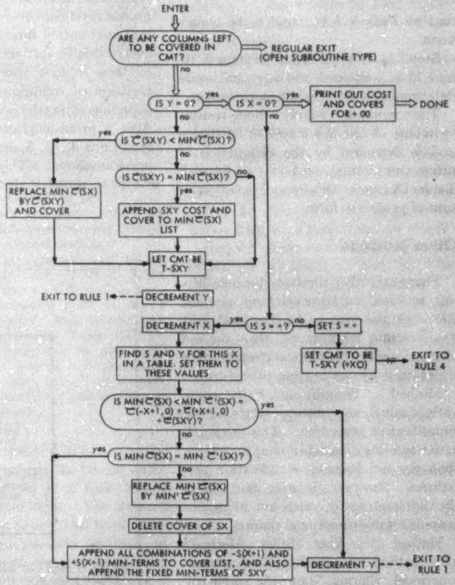

ENTER
ARE ANY COLUMNS LEFT TO BE COVERED IN CMT?
REGULAR EXIT (OPEN SUBROUTINE TYPE)
no
IS Y = 0?
yes
IS X = 0?
yes
PRINT OUT COST AND COVERS FOR + 00
DONE
no
no
IS 𝒞(SXY) < MIN 𝒞(SX)?
yes
REPLACE MIN 𝒞(SX) BY 𝒞(SXY) AND COVER
no
IS 𝒞(SXY) = MIN 𝒞(SX)?
no
yes
APPEND SXY COST AND COVER TO MIN 𝒞(SX) LIST
LET CMT BE T-SXY
EXIT TO RULE 1
DECREMENT Y
DECREMENT X
yes
IS S = +?
no
SET S = +
SET CMT TO BE T-SXY (+XO)
EXIT TO RULE 4
FIND S AND Y FOR THIS X IN A TABLE. SET THEM TO THESE VALUES
IS MIN 𝒞(SX) < MIN 𝒞'(SX) = 𝒞(-X+1,0) + 𝒞(+X+1,0) + 𝒞(SXY)?
yes
no
IS MIN 𝒞(SX) = MIN 𝒞'(SX)?
yes
no
REPLACE MIN 𝒞(SX) BY MIN' 𝒞(SX)
DELETE COVER OF SX
APPEND ALL COMBINATIONS OF -S(X+1) AND +S(X+1) MIN-TERMS TO COVER LIST, AND ALSO APPEND THE FIXED MIN-TERMS OF SXY
DECREMENT Y
EXIT TO RULE 1

Fig. 6 (right). Symbolic routing decision flow chart

Appendix

An ordering for applying the five rules is shown in Fig. 4. The Routing Decision Subroutine (R. D. Sub.) is shown in Fig. 5. A more detailed symbolic flow chart for R. D. Sub. is given in Fig. 6. In Fig. 6 the branch naming code *SXY* is used to facilitate branching. In this code, Y is the number of incomplete Rule 5 branches left on Rule 4 branch number X. Sign S is $-$ if the CMT is the first half of Rule 4 branch number X, or S is $+$ if the CMT is the final half of Rule 4 branch number X. Whenever a branch is completed (i.e., a possible cover is obtained), a new SXY is formed to indicate which incomplete branch should be covered next, and its minimization table, $T\text{-}SXY$, becomes the new CMT.

These flow charts are meant primarily as a guide for coding this method for a digital computer, rather than as a practical organization for hand computation. In performing this method by hand, it may be apparent that a different rule is more appropriate at some stage of the minimization procedure than the rule which the flow chart specifies. Until branching is needed, R. D. Sub. will follow the double-lined paths exclusively.

References

1. BOOLEAN MINIMIZATION, E. J. McCluskey. *Bell System Technical Journal*, New York, N. Y., vol. 35, no. 6, Nov. 1956, pp. 1417–44.

2. SWITCHING CIRCUITS AND LOGICAL DESIGN, S. H. Caldwell. John Wiley & Sons, Inc., New York, N. Y., pp. 145–58.

3. ALGEBRAIC TOPOLOGICAL METHODS FOR THE SYNTHESIS OF SWITCHING SYSTEMS. PART III: MINIMIZATION OF NONSINGULAR BOOLEAN TREES, J. P. Roth, E. G. Wagner. *IBM Journal of Research and Development*, New York, N. Y., vol. 3, no. 4, Oct. 1959, pp. 326–44.

4. AN ALGEBRAIC THEORY FOR USE IN DIGITAL COMPUTER DESIGN, E. C. Nelson. *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., vol. EC-3, no. 3, Sept. 1954, pp. 12–21.

5. APPLICATION OF BOOLEAN ALGEBRA TO SWITCHING CIRCUIT DESIGN AND TO ERROR DETECTION, D. E. Muller. *Ibid.*, pp. 6–11.

6. THE MINIMALITY OF RECTIFIER NETS WITH MULTIPLE OUTPUTS INCOMPLETELY SPECIFIED, R. McNaughton, B. Mitchell. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 264, no. 6, Dec. 1957, pp. 475–80.

An Algorithm for Determining Minimal Normal Forms of an Incomplete Truth Function

THOMAS H. MOTT, JR.
MEMBER AIEE

IN certain applications of truth-function logic it is of interest to consider the minimization problem in a more general form.[1] So-called "algebraic" procedures for determining the irredundant forms of a truth function fall into two categories depending on whether the method requires preliminary expansion of a formula into developed normal form or not. The methods of McCluskey[2] and Petrick,[3] for example, require canonical expansion of a formula; whereas the table of ratio functions by Gazale[4] and a method developed by the author[5] determine irredundant forms solely from the list of prime implicants themselves. The latter (referred to hereafter as the "method of iterated consensus") arrives at minimal forms by an iterative scheme where the rule of consensus is applied repeatedly to the prime implicants. The purpose of this paper is to indicate how algebraic methods of the second variety apply to the more general case where the class of simplest normal truth functions to be determined is that of formulas equivalent to a given formula Φ under the hypothesis that certain conjunctions of letters of Φ are always false.

Synthesis of Switching Circuits Having 'Don't Care' Conditions

In switching network design this corresponds to the problem of synthesizing a circuit using fewest components where, for one reason or another, the designer "does not care" about the existence or nonexistence of an output signal for certain combinations of input signals. Thus, suppose a function of four variables $A_1, \ldots, A_4$, where $A_1, \ldots, A_4$ represent the code for an excess-3 decimal digit. In actuality only the following combinations ever occur as input signals:

A_1	A_2	A_3	A_4	Decimal Equivalent
0	0	1	1	0
0	1	0	0	1
0	1	0	1	2
0	1	1	0	3
0	1	1	1	4
1	0	0	0	5
1	0	0	1	6
1	0	1	0	7
1	0	1	1	8
1	1	0	0	9

Therefore, the responses of the network to the remaining combinations may be disregarded.

A_1	A_2	A_3	A_4
0	0	0	0
0	0	0	1
0	0	1	0
1	1	0	1
1	1	1	0
1	1	1	1

Consequently, a switching circuit which accepts only excess-3 decimal digits as input signals need not have its outputs specified for the second set of combinations because the circuit will never be given those as legitimate inputs. The truth function corresponding to the output of such a circuit is called an "incomplete" truth function. It is a function whose value is specified for some combinations of values of its variables, but not for all. It may be viewed as denoting a class of 2^k completely specified truth functions, where the conjunctions of letters for whose value the function is undefined have k distinct completions, thus giving rise to 2^k different possible truth functions. The problem is that of finding the set of irredundant normal formulas that are equivalent, in a sense to be explained later, to the class of such functions. When the class has only one member, that is, when the function to be simplified is without "don't care" conditions, the problem reduces to the simplification problem posed originally by Quine.[6]

Theoretical Discussion

Let H be a truth-functional formula which says that certain conjunctions of letters of Φ are false; then a weak equivalent of Φ is a normal formula Ψ such that the formula $H \supset (\Phi \equiv \Psi)$ is valid. φ is a weak prime implicant of Φ if, and only if, φ implies a weak equivalent Ψ of Φ and subsumes no shorter clause which implies a weak equivalent of Φ. (A clause subsumes another clause if, and only if, all the letters whose conjunction is the latter clause are among those whose

---

Paper 59-1123, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 11–16, 1959. Manuscript submitted June 1, 1959; made available for printing September 21, 1960.

THOMAS H. MOTT, JR., formerly with RCA Laboratories, Princeton, N. J., is now with Lockheed Electronics Company, Bedminster, N. J.

The author wishes to thank E. W. Veitch and J. T. Chu for their helpful comments.

conjunction is the former.) A formula is said to be weakly irredundant when no well-formed part of it can be eliminated without sacrificing weak equivalence.

The validity of the formula

$$\bar{D}_1.\bar{D}_2\ldots\supset(\Phi\equiv\Phi\vee D_1\vee D_2\vee\ldots)\qquad(1)$$

where Φ stands for an incomplete truth function whose "don't care" conditions D_1, D_2, ... are all false for all values of the constituent literals, justifies treating as weak equivalents of Φ the 2^k members of the class of truth functions resulting from assigning truth values in all possible ways to the k completions of D_1, D_2, ... for $k=1, 2, \ldots$.

Theorem 1

Any weakly irredundant normal weak equivalent of Φ is a disjunction of prime implicants of $\Phi\vee D_1\vee D_2\vee\ldots$.

Proof. Every clause ψ of a weak normal equivalent Ψ of Φ implies Ψ and under the hypothesis implies $\Phi\vee D_1\vee D_2\vee\ldots$. If ψ is not a prime implicant of $\Phi\vee D_1\vee D_2\vee\ldots$, then ψ subsumes a shorter formula ψ' which implies $\Phi\vee D_1\vee D_2\vee\ldots$ and, therefore, under the hypothesis, Ψ. Now to say that $\psi\vee\Psi'$ is equivalent to $\psi'\vee\Psi'$, where ψ subsumes ψ' and Ψ' is the rest of the normal formula Ψ without ψ, is the same as saying that ψ' implies $\psi\vee\Psi'$. Consequently, Ψ has one or more redundant occurrences of literals, in ψ, which could be deleted without sacrificing weak equivalence; hence Ψ is not a weakly irredundant normal equivalent.

Theorem 1, which is an extension of Quine's basic theorem,[6] justifies obtaining the list of weak prime implicants of Φ when the weak simplest forms of Φ are wanted. By the definition of a weak prime implicant, such a list is obtained whenever we obtain the list of prime implicants of the function derived from Φ by assigning values to D_1, D_2, ... which are all true.

Theorem 2

If Ψ is a weakly irredundant normal weak equivalent of Φ, then each clause of the developed normal equivalent of Φ subsumes a clause of Ψ.

Proof. Ψ contains no letters foreign to the developed normal form of Φ, nor, therefore, to any clause φ thereof. Hence any clause of Ψ which φ does not subsume must contain a literal which is negated in φ, and hence the assignment of truth values to literals which make φ true will make all clauses of Ψ false except those which subsumes. But φ implies Ψ since φ implies Φ and Φ is equivalent under the hypothesis to Ψ. Hence φ subsumes a clause of Ψ.

From theorems 1 and 2 and the definition of weakly irredundant, there follows the corollary given as theorem 3.

Theorem 3

Any weakly irredundant normal weak equivalent Ψ of Φ is a disjunction of prime implicants of $\Phi\vee D_1\vee D_2\vee\ldots$ such that every clause of the developed normal form of Φ subsumes a clause of Ψ, and if any clause of Ψ is eliminated there is at least one clause of the developed normal form of Φ which does not subsume a clause of Ψ.

A strong prime implicant of Φ is a prime implicant of the truth function that obtains when the "don't care" conditions of Φ are assigned values which are all false for all values of the constituent literals. Every strong prime implicant of Φ subsumes one or more weak prime implicants of Φ, but not every weak prime implicant of Φ is subsumed by a strong prime implicant. (In certain instances a strong prime implicant is also a weak prime implicant.) Where Ψ and Ψ' are normal weak equivalents of Φ, it is said that Ψ subsumes Ψ' if it is the case that all the weak prime implicants whose disjunction is Ψ' are among the weak prime implicants whose disjunction is Ψ. For example, $\psi_1\vee\psi_3\vee\psi_6$ will be said to subsume $\psi_3\vee\psi_6$, provided that $\psi_3\vee\psi_6$ itself is a normal weak equivalent of Φ. In view of the fact that the formula

$$[H\supset(\psi\vee\Psi\equiv\Psi)]\equiv[H\supset(\psi\supset\Psi)]\qquad(2)$$

is valid, where Ψ is a normal weak equivalent of Φ, ψ a weak prime implicant, and H the hypothesis that the "don't care" conditions of Φ are all false for all values of the constituent literals, one is justified in dropping subsuming normal weak equivalents in the quest for weakly irredundant normal weak equivalents of Φ.

Theorem 4

Given any strongly irredundant normal strong equivalent of Φ, the result of replacing every strong prime implicant φ of Φ occurring therein by a weak prime implicant subsumed by φ—and dropping duplicate occurrences, if any, of weak prime implicants that are subsumed by two or more such φ's—is a weakly irredundant normal weak equivalent of Φ provided that it subsumes no other normal weak equivalent obtained in an analogous manner.

Proof. Let Σ be any strongly irredundant normal strong equivalent of Φ. Every clause of the developed normal form of Φ subsumes a clause of Σ. Suppose φ subsumes a weak prime implicant ψ of Φ, where φ is any clause of Σ. (φ and ψ may be identical clauses.) Every clause of the developed normal form of Φ which subsumes φ also subsumes ψ. Let Ψ be the expression resulting from replacing each clause of Σ by a weak prime implicant of Φ in accordance with the theorem. Every Ψ obtained in this way is a normal weak equivalent of Φ. Moreover, every clause of the developed normal form of Φ subsumes a clause of Ψ. By the definition of a subsuming normal weak equivalent, if Ψ subsumes no other normal weak equivalent of Φ, every clause of Ψ is such that none implies a disjunction of the remaining clauses of Ψ. Thus, for any clause, ψ, of Ψ there is a clause of the developed normal form of Φ that subsumes ψ but does not subsume any other clause of Ψ. Hence if any clause is eliminated from Ψ, say ψ, there is a clause of the developed normal form of Φ which subsumes ψ but does not subsume a clause of Ψ. By theorem 3, Ψ is a weakly irredundant normal weak equivalent of Φ.

Method for Determining Irredundant Weak Equivalents

Theorems 1 and 4 make it possible, given an incomplete truth-function formula Φ and the list of its "don't care" conditions D_1, D_2, ..., to arrive at weakly irredundant normal weak equivalents, as follows. Applying either Gazale's method or the method of iterated consensus to the prime implicants of the function obtained from Φ when D_1, D_2, ... are assigned values that are all false for all values of the constituent literals, the strongly irredundant normal strong equivalents of Φ are first obtained. Next is determined the prime implicants of the truth function that derives from Φ when D_1, D_2, ... are assigned values which are all true for all values of the constituent literals. (These are the weak prime implicants of Φ.) Then, for any given strongly irredundant normal strong equivalent of Φ, normal weak equivalents are formed that result from replacing every occurrence of the strong prime implicants occurring therein by an occurrence of some weak prime implicant that is subsumed by the strong prime implicant in question, for all such weak prime implicants. This is done for every strongly irredundant normal strong equivalent of Φ. From the list of normal weak equivalents thus obtained, any normal weak equivalent that subsumes another is deleted. The remaining ones then constitute weakly irredundant normal weak equivalents of Φ.

Example 1

To illustrate, let

$$a\bar{b}cd \vee \bar{a}bc\bar{d} \qquad (3)$$

be the incomplete truth function Φ, and

$$\bar{a}bd,\ \bar{a}\bar{b}\bar{d},\ ac\bar{d},\ \bar{b}c\bar{d},\ \bar{a}\bar{c} \qquad (4)$$

the set of "don't care" conditions D_1, D_2, $\ldots$, D_5. It is desired to determine simplest normal forms of Φ under the hypothesis that the $D_1, D_2, \ldots D_5$ are always false. The procedure is as follows. First, the strongly irredundant normal strong equivalents of Φ are determined. This is done by applying an appropriate algebraic method to the prime implicants of the function that obtains from Φ when the $D_1, D_2, \ldots D_5$ are assigned all false values. In this example there is only one strongly irredundant normal strong equivalent of Φ; namely, the function so obtained itself:

$$a\bar{b}cd \vee \bar{a}bc\bar{d} \qquad (5)$$

Next, the weak prime implicants of Φ are determined, or, in other words, the prime implicants of the function that obtains from Φ when the D_1, D_2, $\ldots$, D_5 are assigned values which are all true. They are

$$\bar{a}b,\ a\bar{b}c,\ \bar{a}\bar{c},\ \bar{a}\bar{d},\ c\bar{d} \qquad (6)$$

Finally, normal weak equivalents of Φ are formed that result from replacing in equation 5 every occurrence of a strong prime implicant by a weak prime implicant that it subsumes in equation 6, for all such weak prime implicants, as follows:

$$a\bar{b}c \vee \bar{a}b$$

$$a\bar{b}c \vee \bar{a}\bar{d}$$

$$a\bar{b}c \vee c\bar{d} \qquad (7)$$

Since none of the normal weak equivalents subsume another in equation 7, they constitute weakly irredundant normal weak equivalents of the incomplete truth function $a\bar{b}cd \vee \bar{a}bc\bar{d}$ whose "don't care" conditions are $\bar{a}bd$, $\bar{a}\bar{b}\bar{d}$, $ac\bar{d}$, $\bar{b}c\bar{d}$, and $\bar{a}\bar{c}$. Of these, the ones corresponding to simplest normal forms may be selected according to whatever criterion of simplicity is used.

Example 2

Consider the incomplete truth function

$$abc\bar{d} \vee a\bar{b}cd \vee \bar{a}\bar{b}\bar{c}d \vee \bar{a}bc \qquad (8)$$

whose "don't care" conditions are

$$\bar{a}b\bar{c},\ \bar{a}\bar{c}d,\ \bar{b}c\bar{d} \qquad (9)$$

As in the case of equation 3, there is only one strongly irredundant normal strong equivalent of equation 8; namely,

$$a\bar{b}cd \vee \bar{a}\bar{b}\bar{c}d \vee \bar{a}bc \vee bc\bar{d} \qquad (10)$$

Next, the weak prime implicants of equation 8 are determined as follows:

$$a\bar{b}c,\ \bar{a}b,\ \bar{a}\bar{c},\ \bar{a}\bar{d},\ c\bar{d} \qquad (11)$$

Of the strong prime implicants occurring in equation 10 only one, $\bar{a}\bar{b}\bar{c}d$, subsumes more than one weak prime implicant; namely, $\bar{a}\bar{c}$ and $\bar{a}\bar{d}$. Hence equation 10 gives rise to only two normal weak equivalents of equation 8, neither of which subsumes one another:

$$a\bar{b}c \vee \bar{a}\bar{c} \vee \bar{a}b \vee c\bar{d}$$

$$a\bar{b}c \vee \bar{a}\bar{d} \vee \bar{a}b \vee c\bar{d} \qquad (12)$$

Equation 12 therefore represents two weakly irredundant normal weak equivalents of equations 8 and 9.

Example 3

As a final example consider the incomplete truth function

$$\bar{a}bc\bar{d} \vee \bar{a}\bar{b}\bar{c}d \qquad (13)$$

whose "don't care" conditions are said to be

$$a\bar{b}c,\ \bar{a}b\bar{c},\ \bar{a}bd,\ ac\bar{d},\ \bar{a}\bar{c}d,\ \bar{b}c\bar{d} \qquad (14)$$

First, all the strongly irredundant normal strong equivalents of equation 13 are determined by applying Gazale's method or the method of iterated consensus to the prime implicants of the truth function that obtains from equation 13 when D_1, D_2, $\ldots$, D_6 are all assigned values which are false. As in equations 3 and 8 there is only one strongly irredundant normal strong equivalent of equation 13; namely,

$$\bar{a}bc\bar{d} \vee \bar{a}\bar{b}\bar{c}d \qquad (15)$$

Next, the weak prime implicants of equation 13 are determined. They are the prime implicants of the function deriving from equation 13 when $D_1, D_2, \ldots, D_6$ are given values which are all true, and are

$$\bar{a}b,\ a\bar{b}c,\ \bar{a}\bar{c},\ \bar{a}\bar{d},\ c\bar{d} \qquad (16)$$

Of the weak prime implicants, $\bar{a}bc\bar{d}$ subsumes $\bar{a}b$, $\bar{a}\bar{d}$, and $c\bar{d}$, while $\bar{a}\bar{b}\bar{c}d$ subsumes $\bar{a}\bar{c}$ and $\bar{a}\bar{d}$. Hence, replacing in equation 15 each occurrence of $\bar{a}bc\bar{d}$ and $\bar{a}\bar{b}\bar{c}d$ by an occurrence of a weak prime implicant that is subsumed in equation 16, for all such weak prime implicants, and dropping duplicate occurrences of weak prime implicants as they occur, yields the following normal weak equivalents of equation 13:

$$\bar{a}b \vee \bar{a}\bar{c}$$

$$\bar{a}\bar{d} \vee \bar{a}\bar{c}$$

$$c\bar{d} \vee \bar{a}\bar{c}$$

$$\bar{a}b \vee \bar{a}\bar{d}$$

$$\bar{a}\bar{d}$$

$$c\bar{d} \vee \bar{a}\bar{d} \qquad (17)$$

Deleting from equation 17 any normal weak equivalent that subsumes another, leaves

$$\bar{a}b \vee \bar{a}\bar{c}$$

$$c\bar{d} \vee \bar{a}\bar{c}$$

$$\bar{a}\bar{d} \qquad (18)$$

Equation 18, therefore, represents weakly irredundant normal weak equivalents of the incomplete truth function $\bar{a}bc\bar{d} \vee \bar{a}\bar{b}\bar{c}d$ whose "don't care" conditions are $a\bar{b}c$, $\bar{a}b\bar{c}$, $\bar{a}bd$, $ac\bar{d}$, $\bar{a}\bar{c}d$, and $\bar{b}c\bar{d}$.

The Nelson Simplification Procedure

In its approach the present method resembles one that is due to R. J. Nelson.[7] However, between the two methods there are profound differences, as will be illustrated. Nelson's method is, essentially, to find all the strongly irredundant normal strong equivalents of the truth function that obtains from Φ when the "don't care" conditions are assigned values which are all true for all values of the constituent literals. Then if the prime implicants of the disjunction of the "don't care" conditions are among the disjuncts of any of these, they are eliminated and the results include weakly irredundant normal weak equivalents of Φ.

Applying Nelson's method to the first example, equations 3–7, it is found that there is only one strongly irredundant normal strong equivalent of $\Phi \vee D_1 \vee D_2 \vee D_3 \vee D_4 \vee D_5$; namely,

$$a\bar{b}c \vee \bar{a}b \vee \bar{a}\bar{c} \vee c\bar{d} \qquad (19)$$

Of the prime implicant of $D_1 \vee D_2 \vee \ldots \vee D_5$, only one, $\bar{a}\bar{c}$, appears as a disjunct of equation 10. (For convenience the "don't care" conditions in each example have been chosen so that they represent all the prime implicants of $D_1 \vee D_2 \vee \ldots$.) According to Nelson's method, if $\bar{a}\bar{c}$ is eliminated from equation 19 the result includes weakly irredundant normal weak equivalents of Φ, which in this case is simply

$$a\bar{b}c \vee \bar{a}b \vee c\bar{d} \qquad (20)$$

However, under the hypothesis that the "don't care" conditions of equation 4 are all false, it is easy to show that both

$$a\bar{b}c \vee c\bar{d}$$

$$a\bar{b}c \vee \bar{a}b \qquad (21)$$

are equivalents of equation 3. But equation 20 subsumes equation 21. Consequently, $a\bar{b}c \vee \bar{a}b \vee c\bar{d}$ cannot be a weakly irredundant normal weak equivalent of Φ since either $\bar{a}b$ or $c\bar{d}$ can be eliminated without sacrificing weak equivalence.

Applying the method to the third example, equations 13–18, yields

$$\bar{a}b \vee \overline{ac} \vee c\bar{d} \tag{22}$$

as the only "weakly irredundant" normal weak equivalent of equation 13. However, one can show that both $\bar{a}\bar{c} \vee c\bar{d}$ and $\bar{a}b \vee \bar{a}\bar{c}$ are, under the hypothesis that the "don't care" conditions of equation 14 are false, equivalent to equation 13. Since it subsumes these equations, equation 22 obviously cannot be a weakly irredundant weak equivalent of equation 13.

The previous examples serve to show that Nelson's method is not always satisfactory in locating weakly irredundant normal weak equivalents of Φ. The reason for this lies in the manner in which it sets about to select irredundant disjunctions of weak prime implicants. It does so by determining all the strongly irredundant normal strong equivalents of $\Phi \vee D_1 \vee D_2 \vee \ldots$ and then eliminating disjuncts that are not subsumed by a clause of the developed normal form of Φ. It is, of course, true that the resulting expressions are always as simple as the disjunctions of weak prime implicants that correspond to the irredundant normal forms of $\Phi \vee D_1 \vee D_2 \vee \ldots$. But the latter are irredundant only with respect to the clauses of the developed normal form of $\Phi \vee D_1 \vee D_2 \vee \ldots$, and there is no guarantee that the resulting expressions will be irredundant with respect to the clauses of the developed normal form of Φ itself, which is what is required.

Machine Determination of Irredundant Forms

It is of interest to note that the method presented here determines weakly irredundant normal forms of a truth-functional formula Φ without the necessity for expansion of the formula into developed normal form as a preliminary step, as is the case with the method described by McCluskey.[2] Rather, for the present method it suffices that Φ be written simply as a disjunctive normal formula and the "don't care" conditions as conjunctions of different literals occurring in Φ. From the standpoint of programming economy, then, the method is ideally suitable for use with procedures that determine irredundant forms solely from the list of prime implicants themselves, since a method requiring expansion of a formula into developed normal form defeats one of the purposes of such procedures. For this reason either a method like the "table of ratio functions" by Gazale, or the method of iterated consensus of the prime implicants, is used to good advantage in obtaining weakly irredundant forms by means of the present technique.

References

1. AN ALGORITHM FOR DETERMINING MINIMAL REPRESENTATIONS OF A LOGIC FUNCTION, **B. Harris.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., vol. EC-6, June 1958, pp. 103–08.

2. MINIMIZATION OF BOOLEAN FUNCTIONS, **E. J. McCluskey.** *Bell System Technical Journal*, New York, N. Y., vol. 35, 1956, pp. 1417–44.

3. A DIRECT DETERMINATION OF THE IRREDUNDANT FORMS OF A BOOLEAN FUNCTION FROM THE SET OF PRIME IMPLICANTS, **S. R. Petrick.** *Technical Report 56-101*, Air Force Cambridge Research Center, Cambridge, Mass., Apr., 1956.

4. IRREDUNDANT DISJUNCTIVE AND CONJUNCTIVE FORMS OF A BOOLEAN FUNCTION, **M. J. Gazale.** *IBM Journal of Research and Development*, New York, N. Y., vol. 1, 1957, pp. 171–76.

5. DETERMINATION OF THE IRREDUNDANT NORMAL FORM OF A TRUTH FUNCTION BY ITERATED CONSENSUS OF THE PRIME IMPLICANTS, **T. H. Mott.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, vol. EC-9, 1960, pp. 245–52.

6. THE PROBLEM OF SIMPLIFYING TRUTH FUNCTIONS, **W. V. Quine.** *American Mathematical Monthly*, Menasha, Wis., vol. 59, 1952, pp. 521–31.

7. WEAK SIMPLEST NORMAL TRUTH FUNCTIONS, **R. J. Nelson.** *Journal of Symbolic Logic*, New Brunswick, N. J., vol. 20, 1955, pp. 232–34.

Automatic Precipitation and Temperature Reporting Over a Very-High-Frequency Radio System

E. J. WARCHOL
NONMEMBER AIEE

M. S. SACHS
NONMEMBER AIEE

H. S. LORENTSON
MEMBER AIEE

DEVELOPMENT of radio and electronic equipment for telemetering weather data from remote inaccessible locations has been greatly accelerated in recent years. In 1959 the Bonneville Power Administration completed the installation of an automatic precipitation- and temperature-reporting system utilizing standard detection devices, and including a unique telemetering and transmission system. This system was installed above the Hungry Horse Reservoir located on the south fork of the Flathead River in Montana. The Hungry Horse Reservoir, which was built by the Bureau of Reclamation in 1953, has a storage capacity of 3,500,000 acre-feet, of which nearly 3,000,000 are usable for power production. The Hungry Horse Project is designed primarily to increase system prime power on the U. S. Columbia River System and to control floods in the Flathead Valley in the vicinity of Kalispell, Mont., and on the main Columbia River farther downstream.

The problem which arose in connection with the efficient operation of the Hungry Horse Project was how to obtain equipment which would reliably measure daily precipitation and temperature information from the inaccessible drainage areas during the winter months and transmit it to the Hungry Horse power plant. The greater portion of the drainage area contributing to the Hungry Horse Reservoir is inaccessible generally during the months of November through April. The Spotted Bear Ranger Station, located about 60 road-miles from the Hungry Horse Dam and above the reservoir, was selected as the station for measuring precipitation which may serve as an index for the runoff from the entire drainage area above the reservoir. The Ranger Station is manned by the U. S. Forest Service during the summer months. During these months, prior to the installation of the automatic radio gage, the weather data at the ranger station were collected daily and transmitted over the Forest Service radio system. In addition, snow course measurement data are being obtained in the vicinity of the reservoir. These measurements, which are taken generally once a month, are inadequate for forecasting purposes on a continuing daily basis.

Data Requirements

Telemetered precipitation and temperature data are urgently needed for more accurate forecasts of the inflow to the Hungry Horse Reservoir. Daily weather information from the Spotted Bear Ranger Station area promises to serve as an index in determining the rate of snow accumulation during periods when the upstream area is isolated and inaccessible. Daily reports on precipitation and temperature from this station also promise to aid in the more accurate forecasting of daily inflow to the reservoir during periods of fall and winter snow melt. The need for forecasting inflows to the reservoir is particularly important because the storage in the reservoir is considered cyclic since there is insufficient inflow to refill the reservoir each year. In order for this project to perform its share in a coordinated Federal system operation, the Hungry Horse Reservoir must be full at the beginning of each draw-down period, which generally begins in late September or early October. The problem faced in the operation of the project, therefore, is to be continually informed of the water resources so that releases can be made for the production of secondary energy without encroaching on prime power and flood-control requirements and still be assured of refilling. In order to make decisions, by which the above interests may be maximized, it is imperative that forecasts of seasonal volume inflows to the reservoir be known early in the fall and on a continuing daily or weekly basis throughout the winter months and high-water periods in the spring. A continuing daily report on precipitation and temperature by which the water resources can be appraised is therefore highly important to the successful operation of the project.

The type of information required and how frequently it should be transmitted was determined. Reports on accumulated precipitation would yield the most useful data, so the information transmitted would be the water content of rain or snow accumulated from a time before the first snowfall and including the air temperature at the time of transmission. The accuracy requirement for the precipitation would be ±0.1 inch, with a range of 36 inches of precipitation in any form. The accuracy of the temperature reporting would be ±1 F (degree Fahrenheit)

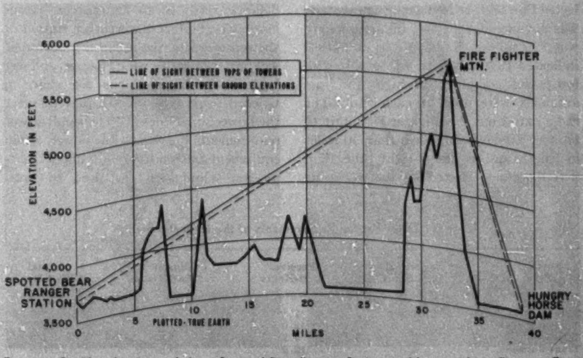


Fig. 1. Profile of radio path from Spotted Bear Ranger Station to Hungry Horse Dam via Fire Fighter Mountain

---

Paper 60-992, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, San Diego, Calif., August 8–12, 1960. Manuscript submitted May 23, 1960; made available for printing October 4, 1960.

E. J. Warchol, M. S. Sachs, and H. S. Lorentson are with the Bonneville Power Administration, Portland, Oreg.

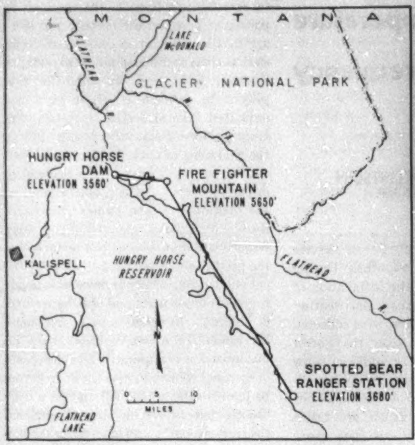


Fig. 2 (left). Vicinity map for Hungry Horse Reservoir weather reporting system

Fig. 3 (above). Originating station equipment house at Spotted Bear Ranger Station in Montana

with a range from −20 F to +100 F. It was also determined that adequate data would be received with transmission twice daily.

Description of System

Very-High-Frequency Radio System

The only feasible communications link to transmit the weather information from the remote station was a point-to-point radio link. The reservoir formed above the dam is relatively crooked and meanders through mountainous terrain on both sides, so a repeater station was necessary. Several repeater sites were investigated with only one site showing any possibility. The radio path from Spotted Bear Ranger Station to the selected Fire Fighter Mountain repeater site was not line-of-sight and the path length was 32.7 miles. Radio checks at 170 mc indicated that this path was usable with low-power equipment if high-gain yagi antennas were used at each end. The radio path from Fire Fighter Mountain to Hungry Horse Dam is less than 10 miles in length and is line-of-sight; the dam was therefore selected as the receiving terminal. Figs. 1 and 2 show the radio paths and path profiles.

The radio equipment is low-power tube type similar to that used for portable single-frequency transceivers. The transmitter output is rated 400 milliwatts, and the receiver sensitivity is rated 0.5 microvolt. Both the transmitters and receivers are equipped for frequency modulation, but this is not used for data transmission. The transmitter carrier is keyed for telemetering and the receiver output utilizes a squelch relay. Frequencies for the two hops are 169.425 mc and 171.825 mc. The receiver and transmitter are isolated from each other at the repeater station by separation and directivity of the antennas in addition to the 2.4-mc frequency separation. All antennas are 4-element yagis.

Since there is no commercial power available at either the weather station or the repeater site, the use of dry batteries, wet batteries with a wind charger, and wet batteries with provisions for charging twice a year were investigated to meet the power supply need. The small power requirements of the low-power radio equipment and relatively infrequent transmissions eliminated the need to overcomplicate the power, since dry batteries for 6 months' operation prove to be of reasonable size. The only real problem with use of dry batteries is the reduction in capacity with temperatures as low as −40 F, which is not uncommon in this area. The obvious solution to this problem is to bury the batteries in watertight aluminum cases far enough underground to prevent the batteries from freezing, which occurs at about −5 F.

Measurement and Telemetering

The equipment obtained for measuring temperature is an oversize bulb-type temperature detector. The temperature bulb drives a transducer directly which converts the measurement to a series of pulses. The pulses consist of two groups of one to ten pulses, with the first group representing the tens digits and the second group representing the units digit. For temperatures below 0 F, the reading starts at 99 for −1 F and becomes progressively smaller for lower temperatures.

The equipment for measuring the accumulated precipitation is a standard weighing-type rain gage with a drum-type recording chart. The precipitation falls into a catch bucket which contains a charge of antifreeze to change snow to water and an oil film to lessen evaporation. A digitizing device is positioned by a motor-driven follow arm which follows the pen arm on the recorder. A small magnetic leaf switch operated by a permanent magnet mounted on the recorder arm actuates the motor to drive the follow arm and digitizer. This system requires power only when there is a change in the

Table I. Daily Battery Requirements of the Repeater Station

Receiver Type	Reporting Cycle, Hours	Receiver Watt-Hours	Transmitter Watt-Hours	Total Watt-Hours
Tube*	12	1.1	0.2	1.3
Transistorized†	12	12.0	0.2	12.2
Tube*	1	2.2	2.6	4.8
Transistorized†	1	12.0	2.6	14.0

* Tube-type receiver switched off between signaling periods.

† Transistorized receiver continuously energized.

Fig. 4. Temperature transducer housing

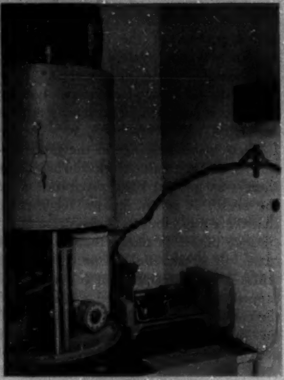


Fig. 5. Precipitation and temperature measuring and transmitting equipment

weight of precipitation in the catch bucket. The digitizer converts the position of the follow arm to three groups of pulses each representing tens, units, and tenths digits.

SEQUENCE OF EVENTS

At the originating station a weight-driven clock determines the interval between reports. Every 12 hours a contact in the clock closes to initiate a sequence of events which make up the transmitted signals. The character of the signals is a series of contact closures which key the carrier of the radio transmitter. The first event that occurs is energization of the transmitter filaments. Shortly thereafter a long pulse is transmitted, followed by the station code designation. After this the three digits of precipitation information are transmitted followed by the two digits for temperature. The equipment then resets for the next transmission 12 hours later. The series of events for a report last approximately 2 minutes with an average carrier on time of about 15 seconds. This shows that the transmit time per day is very small, keeping the battery requirements to a minimum.

Maintaining synchronization of the repeater station where the receiver must be switched off between transmissions is usually a problem. The supplier of this equipment developed a clock system which is very effective in maintaining synchronism. A small battery-operated clock is used at the repeater which drives a disk approximately one revolution in 12.25 hours. The disk has a notch that is equivalent to 0.5 hour in time. When the disk revolves to the leading edge of the notch in about 11.75 hours, the receiver filaments are energized. On receipt of the signal from the originating station, the disk is released from the drive and, being slightly out of balance, it will turn until it reaches the trailing edge of the notch. The disk begins its rotation at the trailing edge of the notch after each transmission from the originating station, so the received signals synchronize the repeater on each report.

Switching the receiver at a repeater station between transmitting periods is reliable, and not only decreases battery requirements but also reduces the equipment operating time to about 2% of continuous receiver operation. Practical battery requirements could be obtained with a low-drain transistorized receiver that is continuously energized; but, as shown in Table I, the switched-tube-type receiver offers considerable battery savings over a continuously energized transistor receiver.

Receiver output at the recording station is fed into a recorder, which indicates the pulses received on a paper tape that is 1 inch wide. In addition to the received pulses, a data time stamp prints time data for the tape for each recording cycle. The recorder is started by the long pulse which precedes any information in the signaling cycle, and it runs long enough to record the complete cycle. The information is relayed daily from Hungry Horse Dam to the dispatching center at Portland, Oreg., and to other interested agencies over power-line carrier and microwave circuits.

HOUSING

A small aluminum building, shown in Fig. 3, was erected to house the equipment at the Spotted Bear Ranger Station. All the equipment is inside this building with the exception of the temperature bulb which is mounted in a louvered shelter near the building that is high enough above the ground to clear the expected snow level as shown in Figs. 4 and 5. As previously mentioned, the batteries are underground near the building to allow additional protection by snow cover. A desiccant tray inside the battery boxes assures dryness.

The housing at the repeater station is merely an aluminum box mounted on a leg of a Forest Service lookout tower. Figs. 6 and 7 show the lookout tower and this housing during installation. The antennas are mounted on legs of the

Fig. 6. Fire Fighter Mountain Forest Service lookout

Fig. 7. Fire Fighter Mountain repeater during construction

tower, and the batteries at the repeater are also underground. Since the snow is blown off the exposed mountain peak at the repeater station site, a snow fence was erected to provide a snow cover over the battery boxes.

Experience

The first winter of operation in 1959 indicated several weak points which must be remedied to assure continuous operation throughout the winter months. The first trouble that developed was that the clock at the repeater apparently stopped during an unusually severe cold period that occurred before there was any snow cover. The mercury batteries that drive the clock apparently dropped in temperature until their internal resistance became too high to supply adequate pulses to drive the clock. A remedy for this problem is to parallel a large capacitor across the battery which will supply the relatively large magnitude required of the short pulses which drive the clock.

When the repeater clock stopped the system lost synchronism, and it took approximately 3 weeks to resynchronize. To eliminate the long synchronizing period it is proposed to use an hourly reporting period instead of the original 12-hour period. If the system loses synchronism on an hourly cycle, it will come back in within 44 hours. Sufficient battery capacity is provided for an hourly reporting period.

There was no change in the precipitation report for about 2 weeks during which there was general light precipitation in the area. It is not known for certain that there was snowfall at the originating station, but it is likely that there was some. It is, therefore, assumed that the orifice of the precipitation collector was capped over by snow or ice. The part of the collector protruding from the roof of the building is painted black and is designed to lessen the chance of bridging over. No remedial measures will be taken this year as the only reliable approach to prevent bridging over is to apply heat around the orifice, which is expensive at a location with no commercial power. Also, it is not known whether or not the orifice bridging is apt to occur frequently or only once in every several years in this location.

Since the equipment was "debugged" and placed in service, there have been two outages attributed to trouble in the radio receiver at the recording station. The first outage was caused by improperly inserting a subminiature tube in its socket during routine maintenance. The next outage was due to an increase in receiver noise which required readjustment of the squelch control to reactivate the circuit.

One other outage of undetermined origin caused the system to lose synchronism.

Summary

Although experience with this type of precipitation and temperature telemetering has been limited, it does indicate that this system will be adequate for this application. Future installations, however, may or may not utilize this type of system because weather data requirements may be expanded, and technological advances in the field may provide more reliable systems for measurement and telemetering of these data. Selection of the best type of system for a specific application depends on the weather data requirements and the types of equipment available.

Digital Telemetering Techniques

A. BROTHMAN
NONMEMBER AIEE

E. H. BROTHMAN
NONMEMBER AIEE

R. D. REISER
NONMEMBER AIEE

DIGITAL telemetering is an information transmission operation in which meter information is processed from source to receiving point in digital form. In the operation, the information is developed in digital form at the source, is transmitted in this form over a long-distance link, and is digitally received and processed at the receiving point. This definition demands, in turn, a definition of digital form and long-distance communications links.

Digital information is composed of alphabetic and/or numeric characters which, in and of themselves or in combination, form words. From a data transmission perspective, characters are in turn formed of binary units called bits, i.e., units of information whose significance is drawn exclusively from their 0 or 1 status in given positions of a standard format or array. Digital information at the character level is distinguished from its opposed form, analog information, in that the spectrum of possible characters at each character position of a word are arrayed at the discrete and separate levels of a staircase.

Long-distance communications links are best considered as information links on which the normal modes of information transfer is by on–off keying of a loop current or a tone, or by shifting between "space" and "mark" tones, or by the modulation of a carrier. The specific characterization "long distance" is here intended to denote the use of long-distance information transfer techniques, rather than to impose a distance requirement on the information path.

Inevitably, any discussion of digital telemetering includes not merely distinguishing it from analog telemetering techniques, but weighing the advantages of one opposed to the other. In analog telemetry, meter information is given a pulse-width or frequency-analog form at the source, is transmitted over the link in terms of pulse-width keying or carrier modulation actions, and is received and translated by analog-to-analog conversion actions. Analog telemetry has the advantage of requiring low-cost equipment and of having, in a sense, a superior capability in trending a meter information source. Because of the additive inefficiencies of the several analog-to-analog translations which are involved, and because of the inevitable error which the link injects in handling pulse-width or variable-frequency analog information, analog telemetry suffers from a large error susceptibility; see Figs. 1 and 2. In contrast, digital telemetering provides greater accuracy at the source than is usual in frequency or pulse-width analog conversions, and once the digital form of the data exists, link-injected error can be virtually totally eliminated. As for the receiving end of digital telemetering systems, errors are possible only in the case of equipment malfunctions. Digital telemetering consequently is favored because of greater intrinsic accuracy, direct suitability to data logging and computer control, and, in many instances, a higher information rate. On the debit side of digital telemetering there is a higher equipment cost, a tendency in some instances to include "process noise" as an additive or subtractive factor to the mean value of the meter information in spot or individual readings, and a greater complexity in imposing alarm and drift monitors at the receiving end of the system.

Technology of Digital Telemetering

Component Aspects of Digital Telemeter Systems

A generalized digital telemetering system is indicated in block form in Fig. 3. It illustrates the component design problems: (1) whether to select optimum transducers for the basic measurements or to tie into existing transducers; (2) selection of a means of digitizing the information; (3) selection of a suitable communications link and its use at its maximum capabilities with regard to information rate and/or information security; (4) selection of data transmission equipment which fully utilizes the link's capabilities; and (5) of matching the data receiving equipment to the terminal display, logging, printing, data processing, etc., equipment at the receiving end of the system.

Transducers and Digitizers, Fig. 4

Some few variables are conventionally measured at the source by digital transducers. Instances are: kilowatt-hours (kw-hr); kilovar-hours (kv-hr), flow by

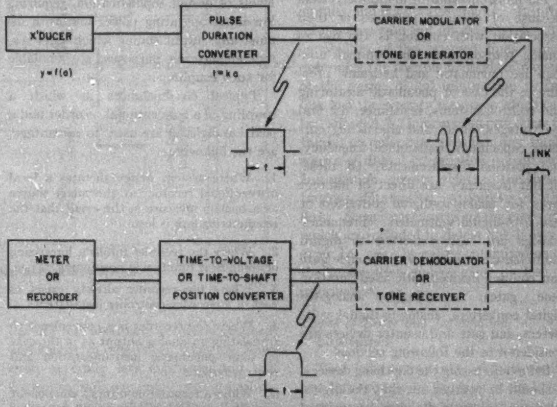

X'DUCER
y = f(a)
PULSE DURATION CONVERTER
t = ka
CARRIER MODULATOR OR TONE GENERATOR
t
t
LINK
METER OR RECORDER
TIME-TO-VOLTAGE OR TIME-TO-SHAFT POSITION CONVERTER
CARRIER DEMODULATOR OR TONE RECEIVER
t

Fig. 1. Pulse-duration analog telemetering. Cascaded error effects:

1. Nonlinearity in pulse-duration converter
2. Distortion of square-wave output of converter in link-terminal equipment and the link
3. Nonlinearity in time-to-voltage or time-to-position converter
4. Error in meter or recorder
5. Error in reading meter or recorder

Paper 61-17, recommended by the AIEE Telemetering Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 2, 1960; made available for printing January 10, 1961.

A. Brothman, E. H. Brothman (student, Barnard College, New York, N. Y.), and R. D. Reiser are with Industrial Process Engineers, Inc., Newark, N. J.

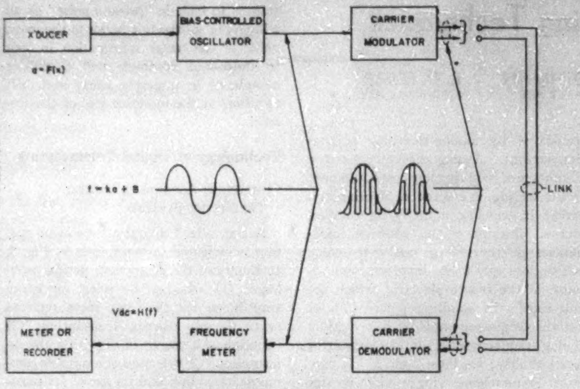


Fig. 2. Frequency analog telemetering. Cascaded effects:

1. Nonlinearity in bias-controlled oscillator
2. Noise modulation of carrier modulator output by the link
3. Lags and errors of frequency meter
4. Error in meter or recorder
5. Error in reading meter or recorder

positive displacement meters; and position. Of these all but position-from-position digitizers or encoders would involve counters as the means of achieving a digital multicharacter telemeter word; and accordingly, counting devices must be considered under the heading of digitizers.

Far more outstanding in number, either because of classical practices or their association with control, is the use of analog transducers. The common analogs use pneumatic and hydraulic pressure in the case of pneumatic measuring and control systems, resistance, d-c and a-c voltages, direct and alternating currents, capacitance, inductance, frequency, and position displacements. Of these, all but frequency are direct or indirect tasks for analog-to-digital converters or analog-to-digital voltmeters. Frequency analogs are best converted to digital multicharacter word form through gate and counter devices with fixed-duration time gates. Accordingly, analog-to-digital converters, analog-to-digital voltmeters, and gate and counter devices are considered in the following sections.

Before discussing the digitizing devices, it should be pointed out that the digital telemeter system's designer has no real reason, aside from personal preference, for entering into the many controversies now raging in the transducer field. It is possible to superimpose digital telemetering on any existing control or instrumentation system.

Recorder-Driven Digitizers

In a frequently encountered method of converting to digital form, the pen drive of a conventional recorder is used to drive a brush-and-commutator type of position encoder. Such a possibility is provided for by many commonly used recorders. Position encoders of varying levels of design sophistication, requiring low-input operating power and with almost any output coding which suits system engineering purposes, are available for such couplings.

Typical circumstances in which a coupling of a conventional recorder and a position digitizer are used to advantage, are the following:

1. Where system design dictates a local conventional recorder at the meter source as a backup measure in the event that the telemetering link is lost.
2. Where the recorder fulfills a linearizing or computing function in positioning its pen.
3. Where the recorder already exists as part of a recorder-controller installation.
4. Where the recorder is a convenience to translating an analog output as in the case of some pneumatic instrumentation and control systems.
5. Where a recorder monitors a component variable of a locally computed telemeter function—or possibly, a computed control function—in which the constituent variables are subject to telemetering.

In part, both the accuracy and the resolution of a recorder-digitizer "couple" are determined by the counting capabilities of the digitizer. In general, an accuracy of conversion of 0.25% to 0.50% of full range is a reasonable expectancy. Couples of this type are low in equipment cost.

Servo-Positioned Brush-and-Commutator Digitizer

Where electrical analogs are encountered, another frequently met digitizing system uses the nulling type of servo positioner to position a brush-and-commutator encoder. In these systems, every possible position of the encoder is uniquely indicated by a feedback analog signal from a transducer which is mechanically coupled to the encoder. In positioning the encoder, a servoamplifier-controlled motor drives the encoder until the feedback signal is nulled against the input process signal. Such a series, shown in Fig. 5, is used to advantage in instances in which low cost is a design parameter or in which the nulling action also fulfills a linearizing or other computing function.

The accuracy and resolution capabilities of this type of digitizer are similar to those in recorder-driven digitizers.

Stepping-Switch Voltage-Staircase-Type Analog-to-Digital Voltmeter

A common digitizer in use for implicit and explicit voltage analogs is the servo-operated stepping-switch voltage-staircase type of digitizer. In principle, the successive throw positions of a stepping switch are placed at the junction points of a voltage divider, and the switch is permitted to step through as much of its tour as is required for its common arm to balance or null a test input. More sophisticated versions of this device use a stepping switch for each digit position, and each switch is permitted to null in sequence according to a "successive approximation principle." Such a device is shown in Fig. 6.

Devices of this type are less costly than electronic digitizers, but are generally slower and have a lower order of reliability due to the stepping-switch components. Sensitivity, resolution, and accuracy can be of any limit ordered by design.

Successive Approximation Logic Analog-to-Digital Voltmeters

Fully electronic successive approximation logic voltmeters are applicable to direct or derived voltage analogs. They have high operating speeds, can provide accuracies and resolutions which are adequate for all system requirements, and have an excellent long-term reliability.

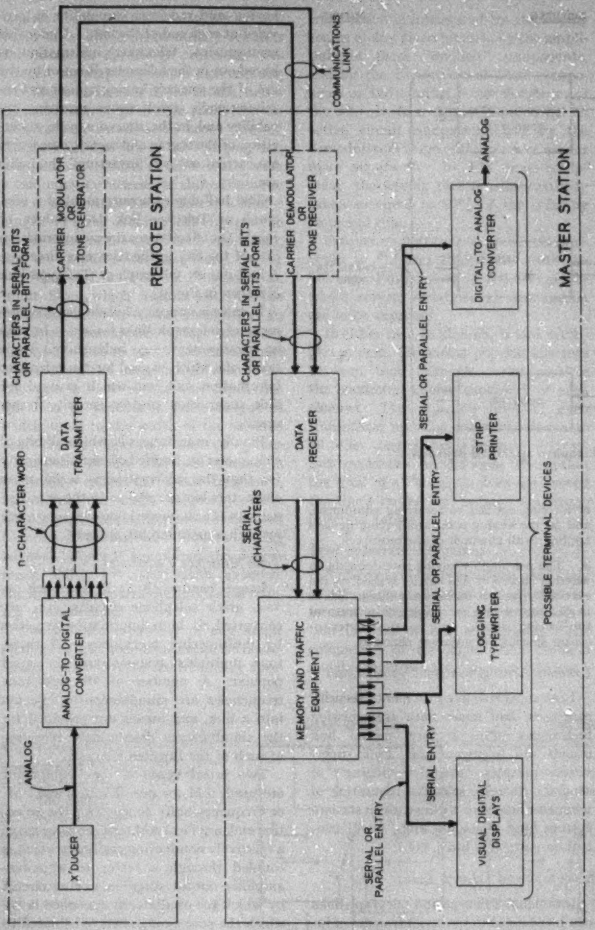

COMMUNICATIONS LINK
REMOTE STATION
X'DUCER
ANALOG
ANALOG-TO-DIGITAL CONVERTER
n-CHARACTER WORD
DATA TRANSMITTER
CHARACTERS IN SERIAL-BITS OR PARALLEL-BITS FORM
CARRIER MODULATOR OR TONE GENERATOR
MASTER STATION
CARRIER DEMODULATOR OR TONE RECEIVER
CHARACTERS IN SERIAL-BITS OR PARALLEL-BITS FORM
DATA RECEIVER
SERIAL CHARACTERS
MEMORY AND TRAFFIC EQUIPMENT
SERIAL OR PARALLEL ENTRY
SERIAL ENTRY
SERIAL OR PARALLEL ENTRY
SERIAL OR PARALLEL ENTRY
SERIAL OR PARALLEL ENTRY
VISUAL DIGITAL DISPLAYS
LOGGING TYPEWRITER
STRIP PRINTER
DIGITAL-TO-ANALOG CONVERTER
ANALOG
POSSIBLE TERMINAL DEVICES

Fig. 3. Generalized digital telemetering system

Their cost is, however, quite high. A typical device of this type is shown in Fig. 7.

Ramp Intersections Type of Analog-to-Digital Converters

Fully electronic intersections type of analog-to-digital converters are used with electrical analog data and enjoy a wide usage. By absolute standards, ramp intersections converters are somewhat inferior to successive approximations logic devices in accuracy, resolution, speed, and long-term reliability. They are, however, generally less expensive and will frequently suffice in many system applications. A device of this type is shown in Fig. 8.

Constant-Width Time-Gate Counters

Constant-width time-gate counters are used to convert frequency analogs to digital form. The time-gate sections of such units can be based on one-shot multivibrator devices, phantastron circuits, or a free-running multivibrator with superimposed gating logic. The counting sections can be based on glow-transfer decade counters, scalers, beam-transfer counters, or flip-flop counters in pure binary or decade-counter arrangements. Little or no preference can logically be stated regarding the time-gate circuitry but a preference for flip-flop counters can be substantiated for counter sections.

Counters

Counters are used with transducers which produce a pulse in equivalence to a given magnitude of the measured information. For this purpose, a wide array of purely mechanical, electromechanical, and electronic counters are available. Much depends on the frequency with which the counted pulses are produced. Generally speaking, electromechanical counting reaches a limit for long-term reliability at roughly 1 pulse per second. Up to this limit, the nonvolatile memory properties and low cost of electromechanical counting are powerful assets and arguments against electronic counting, unless the latter is performed with nonvolatile magnetic devices.

COMMUNICATONS LINKS

Common links for telemetering operations are telegraph, Teletype, and voice-grade telephone circuits, power-line carrier circuits, and microwave circuits. It is the information rate and information density capabilities of the selected links which determine the possible methods of payout of the source digital information to the link and the security measures which can be "afforded." Information rate refers to the bit or pulse rate at which any single source of digital information is paid out. Information density refers to the simultaneous information capabilities of the link. Security measures are the measures which are included to detect link-induced error, i.e., error by noise contamination of the information and/or error by loss of information.

Types of Links and Information Capabilities

The information-rate characteristics of a link are a strict function of the link's band-pass characteristics or the manner in which its bandwidth is assigned to derived channels. In the case of telegraph and Teletype circuits where the link's use is restricted to a single information source at any given time, the information rate in terms of bits per second is at a maximum equal to the band pass of the link. In

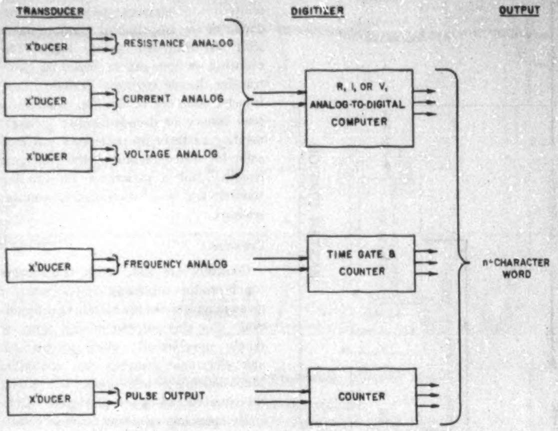


Fig. 4. Types of transducers and required associated digitizers

the case of voice-grade telephone circuits, power-line carrier, and microwave circuits (where the link's bandwidth can be divided into simultaneous derived channels), each channel's information rate is capable of a bit rate which is determined by the width of the derived channel and the keying or modulation properties of the channel-deriving device.

The information density of telegraph and Teletype links is one channel of information at a time. In the case of voice-grade telephone and higher grade telephone circuits, power-line carrier circuits (sometimes called wired wireless circuits), and microwave circuits, an ability to derive simultaneous information channels is developed. The possible number of simultaneous information channels depends on the bandwidth of the link, on the frequency versus attenuation curve of the link, on the signal attenuation which each channel of the channel-deriving equipment permits, on the minimum recommended center-to-center spacing capability of the channel-deriving equipment, and on the expected keying rate to be applied to each channel. In turn:

1. The permissible signal attenuation in each channel depends on whether a frequency-modulated (FM) or amplitude modulated (AM) carrier is used and on the expected signal-to-noise ratio.

2. The center-to-center spacing capability depends on whether an AM or FM carrier is used, on the characteristics of filtering applied to source devices to restrict their dispersion about center-line frequency, on the discrimination and rejection capabilities of filtering applied to receiving equipment, and on the keying rates and keying devices applied to all channels of the multiplex.

3. The permissible keying rate depends on whether an FM or AM carrier is used, on the carrier frequency, on inherent design delays in the source and receiving equipment, on the keying means, and on the center-to-center spacings of adjacent channels.

Common Arrangements of Typical Links

Link-supervision requirements, security measures, and some data transmission techniques often require special link usages and arrangements. Link supervision includes "selective ringing" of separate stations, selective requesting of telemeter functions within a given station, system busy signalling and "busy" control on party-line links, etc.

Telegraph and Teletype Links

Individual Teletype and telegraph links can be adapted to the serial transfer of information bits, to security measures, and to link supervision by arranging the link on a half-duplex operational basis. Here, a mutually exclusive (nonsimultaneous) keying and receiving capability is provided at each end of the link. Under this arrangement, telemeter information is keyed out in the telemeter direction by the use of the source's keying option and receiving end's simultaneous receiving capability and, in the alternate cycle, supervision of the source and security measures are carried out by "inverting" the link's use.

The half-duplex arrangement of a telegraph or Teletype link does reduce, of course, the effective maximum information rate of the link in the telemeter direction. A half-duplex telegraph or Teletype link is shown in Fig. 9.

In some instances, a full-duplex arrangement of telegraph links is used. In such an arrangement, two independent loops exist: one which is used for the telemeter information and one which is used for link supervision and/or security measures.

Finally, in instances in which all bits of a character are handed off simultaneously (so that the information is serial as to characters but parallel as to the constituent bits of a character), parallel loops, one for each constituent bit, are used.

Tone Multiplex Techniques

Large bandwidth facilities, such as voice-grade telephone circuits, etc., are converted to high information densities by multiplexing techniques. Of these, tone multiplex systems are the most popular. A number of discrete tone frequencies are simultaneously injected into a link, and means are provided for the simultaneous discriminate receiving of each of the injected energies.

Two broad types of tone equipment are used: AM, or on–off tones, and FM, or frequency-shift tones. At the injection end, a typical AM tone set consists of: a relatively nondrifting oscillator which is coupled through a buffer to a power-amplifier output stage; a keying circuit by which the oscillator is quenched or by which the power-output stage is biased to and from its relaying function; and sometimes, an output filter by which the output is restricted to the center-line frequency of the oscillator. At the receiving end, a

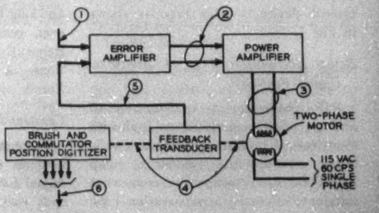


Fig. 5. Servoamplifier-positioned brush-and-commutator digitizer. Signals are:

1—Meter analog
2—Amplified error
3—Control winding
4—Mechanical coupling to feedback transducer and digitizer
5—Feedback
6—Digital output

typical AM set consists of a highly skewed input filter which preferentially restricts the intake to the design center-line frequency, an amplifier which increases the amplitude and power of the intake, and a demodulation stage which yields the direct current envelope of tone "burst." At the injection end of an FM tone set, a typical set consists of an oscillator with a tank circuit which can be shifted between two characteristic stable states; a buffer stage; a keying circuit which engages one or the other of the oscillator's two states; and an output power amplifier. At the receiving end, an FM set consists of an input filter which admits both of the frequency outputs of the transmitter, an amplifier which applies voltage and power gain to the two received frequencies, and a demodulating stage which selectively responds to one of the two input frequencies to produce a pulse output of width equal to the width of the selected tone burst. One of the two tone frequencies of any FM tone channel is commonly called the space frequency, the other the mark frequency. Frequency-shift tone equipment is designated in terms of the nominal minimum spacing between adjacent tones and the space-mark shift. In conservative practice, the shift is one-fourth of the nominal spacing characteristic.

The bit-rate capability of any single channel of AM tone equipment is theoretically fixed solely by the frequency spectrum which is introduced by the keying source at any given bit rate, by the amplitude of these "implicit" components, and by the interaction of these components on the nominal output frequency of the set. In fact, delays imposed by reactive circuit components and by the demodulation stage of the receiver are far more significant. In FM tone equipment, the keying rate is conservatively taken as equal to 80% of the shift by empirical rule.

Frequency-shift tone equipment is preferable to AM tone equipment where the link signal-to-noise ratio is high and/or where severe variations in attenuation are to be expected.

In either case, in planning a tone multiplex system, the widest possible spacings between tone channels make possible the maximum keying capability of each channel. The injection level of each transmitter must be balanced relative to all other tone channels because of cross-talk considerations, and must be such that the total of all channels does not exceed the line's power and voltage amplitude injection rating.

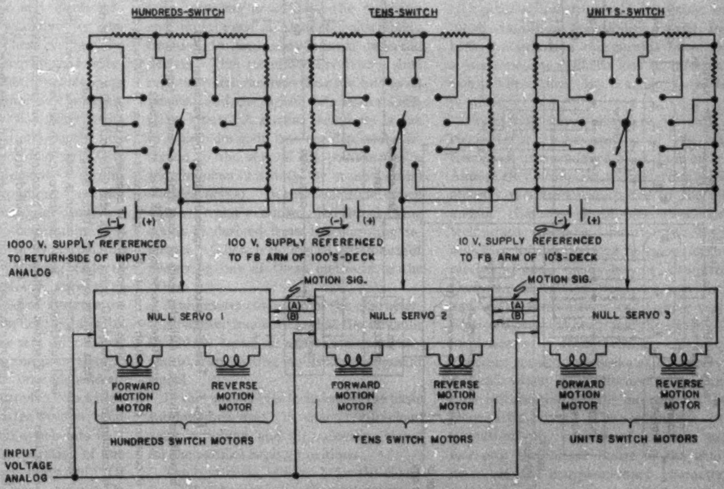
Link Supervision Problems

Regardless of the type of link used and/or the means of achieving selected information densities, a prime consideration is that of link supervision. The complexity of the supervision problem and its details are determined by the overall system design of which the telemeter functions are a part. Much depends on whether the telemeter function is a continuous one, a selectively requested function, an automatically programmed function, whether the system provides for self-initiated telemetering on drift or alarm-limit criteria, whether the telemetering is mixed with remote control, alarm, and control annunciation functions, and whether—in a multistation arrangement—the link is arranged on a radial or party-line basis. There are no pat answers to these problems.

Data Transmission Equipment

The tasks of the data transmission equipment are: (1) to accept a telemeter word from a digitizer and to key the word onto a link at the source end of the system; (2) to receive telemeter words at the receiving office and to hand these off to the appropriate terminal equipment; (3) to exert all necessary security and error-monitoring measures to assure the integrity of the telemeter words against link-injected error.

The keying and receiving functions are built around the decision as to whether each character of the telemeter word is to be handed off in serial-bits or parallel-bits form. The error-monitoring functions are in turn determined by the importance of security to the system purposes. The importance of security to digital telemetering systems which collect data for



1. Forward motion at servo 1 occurs until ΔV is less than 100 volts
2. Forward motion at servo 2 occurs until ΔV is less than 10 volts
3. Forward motion at servo 3 occurs until ΔV is less than 1 volt
4. Servo action is discontinued at any servo until input motion signal from an upstream servo is discontinued
5. Signals A indicate 9-to-0 transit
6. Signals B indicate 0-to-9 transit
7. Each switch wired to hand off its output character by its auxiliary switch deck

Fig. 6. Stepping switch successive approximations analog-to-digital voltmeter

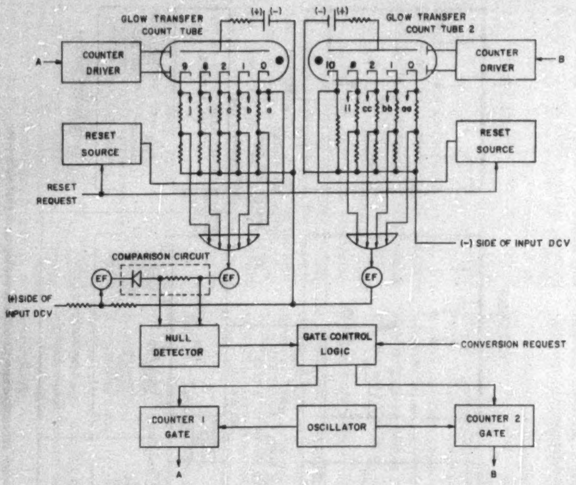

GLOW TRANSFER COUNT TUBE
GLOW TRANSFER COUNT TUBE 2
COUNTER DRIVER
COUNTER DRIVER
A
B
RESET SOURCE
RESET SOURCE
RESET REQUEST
(–) SIDE OF INPUT DCV
COMPARISON CIRCUIT
EF
(+) SIDE OF INPUT DCV
NULL DETECTOR
GATE CONTROL LOGIC
CONVERSION REQUEST
COUNTER 1 GATE
OSCILLATOR
COUNTER 2 GATE

Fig. 7. Two-digit successive approximations logic voltmeter (d-c)

1. By voltage division in cathodes loads of count tubes, a counting action is made to yield a voltage-staircase feed to OR logic at each tube
2. In counting from 0 through 9, count tube 1 applies voltages from 10 to 90 volts d-c in 10-volt steps to unblock diode in comparison circuit
3. In counting from 0 through 10, count tube 2 adds voltages from 0 to 10 volts d-c in 1-volt steps to input analog to reblock comparison circuit
4. After each conversion request, oscillator output is applied by the 1 gate to the 1 counter until null detector "sees" first approximation per item 2
5. After item 4, gate 2 applies oscillator output to counter 2 until second approximation per item 3 is seen by null detector
6. By wiring digital outputs a through j as direct indications of the tens character and outputs aa through ii to indicate complement of beam position, a 2-character digital voltage is obtained. Note that: Count tube 2 is a duodecade (count-of-12) counter; and that: The voltage signals from any cathode of either count tube are purely digital phenomena, being 0 volts d-c when glow beam is absent from the cathode and present at binary 1 level only when glow beam is resident at the cathode

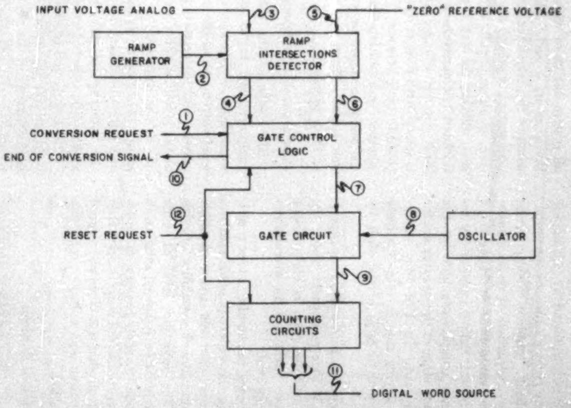

INPUT VOLTAGE ANALOG
"ZERO" REFERENCE VOLTAGE
RAMP GENERATOR
RAMP INTERSECTIONS DETECTOR
CONVERSION REQUEST
END OF CONVERSION SIGNAL
GATE CONTROL LOGIC
RESET REQUEST
GATE CIRCUIT
OSCILLATOR
COUNTING CIRCUITS
DIGITAL WORD SOURCE
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪
⑫

Fig. 8. Ramp-intersections type of analog-to-digital converter

1—Conversion-request signal
2—Continuous linear ramp signal
3—Input analog voltage
4—Ramp to item 3 intersection signal
5—Input reference voltage
6—Ramp to item 5 intersection signal
7—Square-wave gate-control signal. Front porch developed with item 4 and back porch with item 6
8—Digitizing oscillations
9—Gated output consisting of a square-wave item 7 modulation imposed on item 8
10—Digital word output corresponding to item 3
11—Reset signal

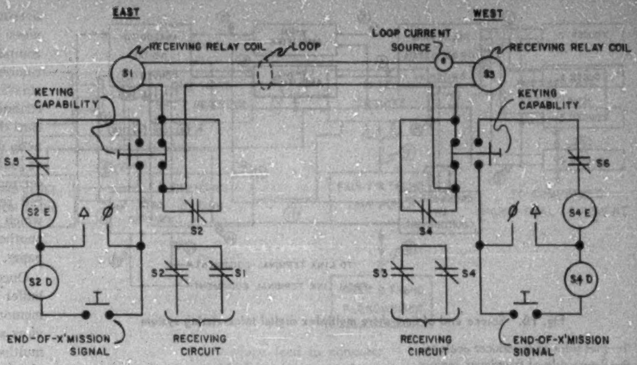


Fig. 9. Half-duplex telegram or Teletype circuit

1. S2 at east end and S4 at west end are mechanically latched relays which are engaged with first keying action of a transmission and are delatched by a signal derived from completion of keying of the last bit
2. S5 at east end and S6 at west end are off-delay timing relays which are engaged by the first received bit of a received transmission and allow sufficient delay for receiving of an incoming transmission

the purposes of a dispatch computer exerting computed control, is obviously greater than its importance where human control is exerted because of the greater intuitive and rational resources of a human being.

Serial-Bits and Parallel-Bits Techniques

The two modes of paying out digital data to a link are: a serial payout of the bits of which each character is composed, or a parallel (simultaneous) payout of such bits. In both instances, the terminal effect is a serial payout as to the characters of the telemeter word. Rarely, if ever, is a parallel payout of characters used, because this would involve enormous bandwidth links for large words.

Both fully electronic and electromechanical devices are available for the serial-bits technique. These program the character bits in a fixed sequence. In electromechanical devices, motorized stepping switches, brush-and-commutator switches, cross-bar switches, and relay ring circuits have been used for programming the bits sequence. In fully electronic serial-bits data transmission equipment, programming is derived from counters operating with superimposed gating logic. In either case, once the basis of operation is a bits sequence, a means of synchronizing the transmitting and receiving ends of the system is required. A variety of techniques are used. Some systems inject the synchronized (synch) data in an adjacent channel; some rely on a synch which is enforced by a common electric power grid; some rely on independent but accurately adjusted "clocking" sources at both ends of the system; and some mix code bits and synch bits in a single channel and use separation of one from the other at the receiving end. Of these techniques, the least dependable are those using independent clocking sources or those using the power grid for synch.

For the character-to-character sequence, either of two methods is employed by the available fully electronic and electromechanical serial-bits devices. In the first, all bits of all characters are presented in parallel to the transmitter, and a pause is developed between adjacent characters to permit the receiving end to hand off the assembled character. In the second method, each character is switched in by a scan switching action. Where repeating of a character is an included function in the event of detected transmission error, the second method is more convenient.

In receiving serial bits, the receiving end must exert a character-assembling function to feed conventional terminal devices. One technique involves a shift register which receives the code bits as its bit-feed and the synch bits as its shift-pulse input. A second technique is one in which the code bits are discretely assigned to the several bit positions of a one-character memory by a sequential clocking action. In this technique, synch data are usually counted, and the clocking action is derived from the counting sequence. There is no logical argument for preferring one of these methods to the other.

A minimum condition of the character-to-character sequence is that the advance should be based on a discrete receiving-end to transmitting-end acknowledgment signal.

In the parallel-bits techniques, the basic subfunctions are: (1) character-to-character switching, and (2) a clocking signal for the parallel keying functions.

The best basis, in item 1, for the clocking of the character-to-character sequence is an acknowledgment of the previously transmitted character. The best basis for item 2 is a fixed time delay derived from the character-to-character switching function. The function of the time delay is to permit switching transiencies to subside before keying is effected.

Transmission Hazards and Error-Monitoring Techniques

To derive the maximum benefits of digital telemetering, cognizance must be taken of the aberrated behavior potential of links. Regardless of the class of the link or its use arrangement, all links are characterized by a constant or base noise, sporadic superimposed noise during periods of link instability, and sporadic high-attenuation intervals. Because of these conditions, possibilities exist for both the insertion of spurious bits into a transmission and the loss of information. These are, in practice, equally possible.

Understandably, therefore, security measures must be applied at the bit, character, or word levels to exclude error. Almost exclusively, in telemetering, such security measures are applied at the bit and/or character level. The common techniques include: parity checking, matching of iterated transmissions at the receiving end, bit-by-bit or complete character checkback from the receiving end to the transmitting end (where verification is carried out), and the use of coincidence or validity checks. Validity checking consists of checks at the receiving end which are performed on redundant information that is inserted by the transmitting end. The redundant information may consist of forming and inserting the mirror image of the character, or the complementary character, or simple format checks. Still another

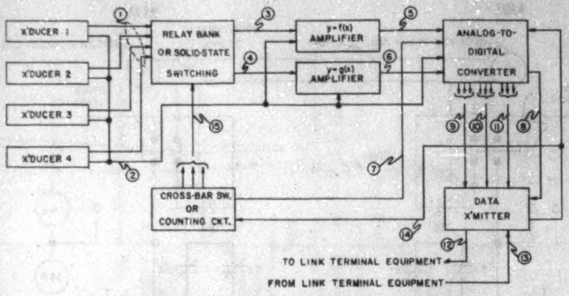

X'DUCER 1
X'DUCER 2
X'DUCER 3
X'DUCER 4
RELAY BANK OR SOLID-STATE SWITCHING
y = f(x) AMPLIFIER
y = g(x) AMPLIFIER
ANALOG-TO-DIGITAL CONVERTER
CROSS-BAR SW. OR COUNTING CKT.
DATA X'MITTER
TO LINK TERMINAL EQUIPMENT
FROM LINK TERMINAL EQUIPMENT
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪
⑫
⑬
⑭
⑮

Fig. 10. Source end of time-share multiplex digital telemetering system

1—Hot sides of transducer analogs
2—Return side of transducer analogs
3—Common pole for all outputs requiring processing by y = f(x) amplifier
4—Common arm for all outputs requiring processing by y = g(x) amplifier
5—Output representing item 3 after linearization and/or zero suppression
6—Output representing item 4 after linearization and/or span correction and/or zero suppression
7—Conversion request signal, preferably with fixed delay from scan switching action, to permit switching transiences in items 3–6 to subside
8—End-of-conversion signal from A-to-D converter applied as transmission request to data transmitter
9–11—Constituent characters of telemeter word
12—Characters to link in serial-bits or parallel-bits form
13—Character-by-character acknowledgment signals from receiving end used to develop character-to-character scan switching in data transmitter
14—End-of-word signal from data transmitter used to reset A-to-D converter and to signal program source to its next program point
15—Program signals to switching center

technique is the use of "self-checking" codes.

For reasons which can be evaluated by a mathematical probability analysis, it goes almost without saying that the cascading of these separate techniques achieves vast improvement in security. Because, in their essence, each one of the techniques is based on redundant information which is superimposed on each character, a powerful cascade can be achieved without an undue accumulation of redundant information if the redundant information is shared by the separate members of the cascade.

In addition to checks on link-induced error, frequently a need exists to monitor the link itself for interruptions of service. Where FM tone equipment is used, link continuity is readily checked by monitoring the space carrier. In instances of direct-current telegraph and Teletype links and in instances in which AM tone equipment is used, link continuity can be monitored by using the ON state of the loop or tone circuit to designate the binary 0 state of information bits. Here, binary 1 bits appear as momentary circuit interruptions of a normally engaged circuit. In some systems, periodic check transmissions and answer-backs are used to check the link and terminal equipment at both ends of the link. Periodic check-transmission measures are frequently accompanied with interlocking whereby a failure of iterated check transmissions automatically invokes standby data storage equipment.

Assaying the relative merits of electromechanical and fully electronic data transmission equipment, a logical preference is for the latter at high information rates and in instances in which a powerful cascade of error-monitoring techniques is used. This preference is based on the bounce and chatter properties of electromechanical devices at high speeds, their changing fly-time characteristics with wear, and the lower space requirements per logic function of electronic circuitry.

Matching Into Receiving-End Terminal Equipment

Data storage devices are sometimes necessary components of digital telemetering systems for a variety of reasons. Some of the reasons are: the intermittent overtaxing of "hard-copy" or display devices, the requirements of coupling to a computer or data processor, or the need to combine or compare data. In such instances, punched paper tape, multichannel magnetic-tape recorders, punch-card devices, magnetic-core arrays, magnetic drums, shift registers, etc., are used as buffer devices between the data-receiving equipment and end devices. A full review of the circumstances under which one of these would be superior to another lies outside the province of this paper.

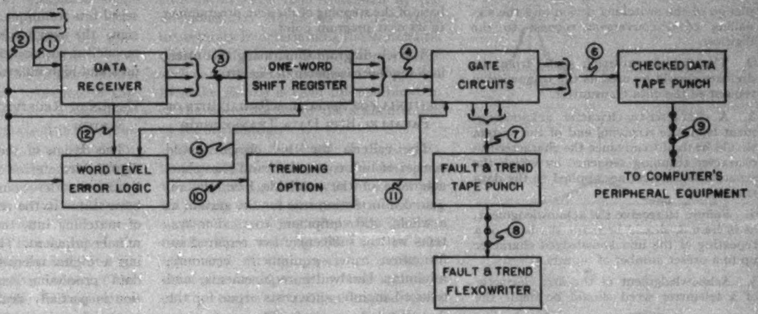
Once information is consigned to any buffer device (and particularly when memory drums and magnetic-core devices are used), the separate sources of a multisource system are not easily or completely and reliably identified by sequence, especially if for any reason a "missed" reading occurs. This raises the question of addressing each piece of information, or at least groups of information. Several techniques are used: addressing at the receiving end on the basis of interlocking with receiving-end request of control over the sources, or addressing on the basis of interlocking with the link equipment, or addressing of the information at the source. Of these, addressing at source is frequently the most reliable technique, even though it constitutes a complication to information-rate and information-density problems. Regardless of where the addressing occurs, a level of protection against the confusing of address characters with data characters by the end devices is eliminated by distinguishing address characters from data characters. By using an extra binary 1 bit in an extra bit position for address characters, such confusion can be avoided.

Addressing is not only a must in loading computers and in entering matrix devices like magnetic-drum and magnetic-core arrays, but it can also be used for traffic purposes in selectively routing to terminal hard-copy devices and visual displays.

Nearly always, telemetered data are logged or visually displayed for observation by personnel. If any rules can be cited at all in regard to hard-copy devices and visual displays, they are:

1. The amount of logged data should be minimized, if at all possible.
2. Log records for the purposes of a later engineering or accounting analysis are best assigned to wide-carriage typewriters.
3. Data which are immediately associated with emergency control measures (such as trending data, alarm printouts, etc.) are most conveniently handled by strip printers.

Fig. 11. Typical arrangement of receiving end of digital telemetering system to link it to computer. Manual trending option also shown



1—Bits data from link

2—Character-by-character acknowledgment signals to link, used also as shift-pulse source for shift register and as drive for a counting function in word-level error logic

3—Character handoffs from data receiver

4—Overpoured characters from shift register

5—Error-marking signals generated at conclusion of loading of a word in the register; used as gating instruction to gate circuits

6—Characters of nonerrored data sentences

7—Characters belonging to words which have been marked for error or trending

8—Tape containing words marked for error or trending; marked as "errored" or "trend" on tape

9—Tape feed to computer

10—End-of-sentence signal derived from counting action applied to item 2

11—Trend-request signals developed by open-gate action on signals (10) by "trend option." Applications of trend request to gate circuits cause word in register to be directed to checked data and fault data tape punches

12—Signal indicating a preset number of repeated transmissions of a character with error on each transmission

Note that: The "word error logic" evaluates completeness of an address applied at source end, examines data for their completeness, and examines address and data characters for nondecodeable bit forms due to source-end equipment malfunctions

4. Data which merely confirm proper operation to a dispatcher should be assigned to visual displays.

5. All printed data should be accompanied by time printouts.

6. Regularly logged data should preferably be typed on preprinted log paper to eliminate confusing data identifications.

7. Means should be included on strip-printer and logging-typewriter records to distinguish normal from alarm data.

Logging typewriters, strip printers, and visual digital displays are available as decimal input devices and as binary input devices. In most instances, a decimal input device with an external fully electronic binary-to-decimal converter is preferable to a binary input device with a mechanical and/or electromechanical binary-to-decimal converter.

In linking data receivers either directly to hard-copy and visual display devices or through buffer devices, provisions should be included to prevent the loading of errored or incomplete data into data logging or data processing equipment. To accomplish this, an ability to request repeats, or an automatic repeating feature, should be built into the telemeter equipment, to be accompanied by an automatic rejection of data at the receiving end which is marked in error by error-monitoring devices. Such provisions can be incorporated at low cost.

System Design Concepts

Initiation of Telemetering Functions

Digital telemetering, like analog telemetering, can be fully continuous telemetering of a single function, selective on-demand telemetering, continuous scan telemetering of a number of functions, automatically engaged telemetering in response to drift or limit detection, repeat-cycle telemetering representing a sweep of scanned functions, or automatically engaged telemetering to assist remote raise–lower control functions.

In any of the multiple-source circumstances cited, other than the one involving drift- or alarm-initiated telemetering, a single analog-to-digital converter which "sees" a switched input is an economical design mode. In cases involving the continuous telemetering of a single source, and in instances of drift- and alarm-initiated telemetering, individual servo-positioned encoders or recorder-encoder couples are the preferred forms of a digitizer.

In drift-monitored functions, a combination of bidirectional counting in application to both directions of motion of a position encoder is a common and eminently satisfactory method of sensing drift from a last-reported value. In such applications, each count which reaches the drift limit initiates telemetering and resets the counter for a new reference. In alarm-limit detection, the position encoder lends itself easily to digital set-point-to-process comparisons through digital logic circuitry.

Where a single analog-to-digital converter is used for a multiplicity of sources, fully electronic units are recommended. In switching electrical analogs into the digitizer, piloted mercury-wetted relays are preferable for low-level analogs, especially if high source impedances are involved. In switching high-level electrical analogs, fully electronic switching is not only an excellent approach wherever it is possible, but for many reasons a highly desirable one. Where mercury-wetted relays are used to switch analog inputs into a digitizer, a cross-bar switch is a more desirable program source than stepping switches or relay ring circuits.

Interlocking of Scanning, Analog-to-Digital Conversion, and Data Transmission

In systems which employ scan switching of a number of sources to a single digitizer, the interlocking of the over-all equipment train is best accomplished by the following:

1. After each switching function, a time delay should be invoked between the com-

pletion of the switching action and the signalling of a conversion request to the digitizer.

2. The end-of-conversion signal from the digitizer should be used as the transmission request to the data transmitter.

3. A character-by-character acknowledgment from the receiving end of the system should be used to advance the character-by-character scanning sequence by which the separate characters are applied to the data transmitter.

4. Failure to receive the acknowledgment, as in item 3, should be made the basis of a repeating of the unacknowledged character up to a preset number of repeats.

5. Acknowledgment of the last character of a telemeter word should be made the basis of the stepping of the scan programmer to its next program point.

A block diagram illustrating this interlock sequence is shown in Fig. 10.

Criteria for Selecting Serial-Bits or Parallel-Bits Data Transmission

The criteria for this choice should consist of link costs, the band-pass characteristic of the available link, the required information rate for the system as a whole, and equipment cost. For systems with a sufficently low required information rate, equipment economy, minimum bandwidth requirements, and reduced maintenance costs argue for the serial-bits technique. With optimum design, the serial-bits technique can be reconciled with maximum security measures and high information rates.

Design of Receiving End of the System

The design of the receiving end of a digital telemetering system is well guided by the considerations which have already been stated in the review of the problem of matching into the receiving and terminal equipment. The problem of matching a digital telemetering system into a data processing or computer installation is partially dealt with in Fig. 11.

An Analysis of Errors in the Calibration of Electric Instruments

F. L. HERMACH
MEMBER AIEE

IT has often been stated that the standards and equipment used to calibrate an electric instrument, or other electric measuring apparatus, should be ten times as accurate as the instrument being tested. This rule-of-thumb usually leads to impossible demands on the accuracy of the standards, particularly if more than two echelons of standards are involved. Troublesome problems should be expected in primary, secondary, and tertiary standardizing laboratories in large organizations. At any echelon, operations on a 10-to-1 basis require standards that become vastly more expensive in first cost and in actual use, because of the type of personnel and time required, than may be warranted by the instruments being tested.

The 10-to-1 rule-of-thumb arises from a natural desire to make certain that errors in the standards used do not affect the accuracy of calibrations. Although this is a desirable goal, it must be remembered that absolute accuracy is not attainable.[1] Thus, a logical procedure is needed to determine what uncertainty or error the standards and the instruments being tested introduce in the measurements. This paper presents such a procedure, and illustrates it with detailed examples. It is based on established principles in the theory of measurements,[2-6] which are applied here to the calibration of measuring apparatus. The application can lead to important advantages in selecting and evaluating standards and in defining and determining the accuracy of calibrations. It is an extension of methods that were originally outlined in an unpublished report by F. B. Silsbee and have been used for many years to evaluate the accuracy of the standards in the Electrical Instruments Section of the National Bureau of Standards.[7]

Calibration is used in this paper in the dual sense of either laying out the scale or determining the accuracy of the indication of the instrument. Standard is used rather loosely, as including reference instruments and comparators, such as potentiometers, as well as standard cells and resistors.

Errors and Uncertainties in Calibrations

The result of the calibration of an instrument, or other measuring apparatus, is often a series of numbers giving the corrections at chosen indications of the instrument. These measured corrections, however, are "uncertain" because of errors in the standards and procedure and in the instrument under test. Some of these errors are largely "systematic," i.e., constant or slowly varying during the measurement. Examples are the difference between the true and nominal values of the resistance of a resistor or of the emf (electromotive force) of a standard cell used in a potentiometer. Others are random errors, which vary from one measurement to the next. Examples are changes in the test instrument or standard caused by fluctuations in temperature, and errors of setting the pointer of the instrument over a division line on the scale. The distinction between systematic and random errors sometimes depends on the length of time involved.[4] For example, the temperature of a standard may be constant or slowly varying during the short time required for the calibration of a single instrument, but may vary in a random way about the average value over a period of one month.

There are also mistakes or blunders, such as misread scales and errors caused by malfunctioning electronic equipment. Their possibility plagues all measurements and can be reduced by increasing the reliability of both the observer and equipment, or as is often done at the National Bureau of Standards, by making duplicate measurements, with different observers and, sometimes, different standards.

Other errors result from incorrect methods and procedures, such as failure to account for lead resistance in the calibration of a millivoltmeter. There is no substitute for the knowledge and

---

Paper 61-63, recommended by the AIEE Indicating and Integrating Instruments Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 27, 1960; made available for printing December 5, 1960.

F. L. Hermach is with the National Bureau of Standards, Washington, D. C.

The author is grateful to F. M. Defandorf for his suggestions and encouragement, to W. J. Youden for his considerable help, and to F. B. Silsbee for permission to use material in his unpublished report.

experience required to eliminate such mistakes.

The random errors determine the precision (repeatability) of the result. In a sense both systematic and random errors determine the accuracy (closeness of the result to the unknown true value), since a value cannot be accurate if it is imprecise, unless it is the average of many measurements.[8] Ordinarily time does not permit more than two or three measurements at any given setting of the instrument under calibration.

The correction to an instrument is the value to be added algebraically to its reading to obtain the true value. (The error is opposite in sign.) The purpose of calibration is to determine the corrections under defined conditions. The uncertainty, U, in the calibration is the estimate of the possible error; i.e., the estimated range of values on each side of the measured correction within which the true correction is believed to lie. These quantities may, of course, be expressed as decimal fractions of the indicated value, and such fractional corrections are small for a useful instrument.

The relationships between these quantities are illustrated in Fig. 1, where V is the variable, such as voltage, being measured. In Fig. 1(A), S is the nominal value of voltage indicated by a constant setting of the dials of a standard, such as a potentiometer. Because of the systematic error, E_s, of the standard, the average value of the applied voltage required to balance the potentiometer differs from S and is shown as T. The uncertainty caused by the random error of the standard is shown as U_s under the bell-shaped curve around T and will be defined more exactly later. The height of this normal error curve at any given voltage is proportional to the probability that a voltage in a small interval including this value will be obtained at a single balance.[8,9]

If a constant voltage is applied to the instrument under test, the readings obtained, as shown in Fig. 1(B), will vary because of the random error of the instrument. The resulting uncertainty is shown as U_c under the bell-shaped curve around the average readings, I. Ordinarily the spread of these readings is considerably larger than U_s. Again the height is the probability of obtaining an indication in a small interval around the abscissa.

Assume now that the instrument is calibrated by setting the dials of the potentiometer, adjusting the voltage to balance the potentiometer, and then reading the instrument under test. The average applied voltage is shown as T in Fig. 1(C), and the average value indicated by the instrument is shown as I. The total uncertainty in determining the systematic error, E_c, of the instrument from a single such calibration (if corrections are applied to make $E_s=0$) is then the combination of the uncertainties of the standard and of the instrument.

Determining the length of the line, E_c, in Fig. 1(C) is in some ways analogous to determining the length of an object by placing it on a graduated meter bar, so that one end coincides with the zero mark, then reading the bar at the other end of the object. A series of observers would vary in their estimates of the tenths of millimeters, so there would be a distribution of estimated lengths of the object. No one would seriously propose using a microscope to align one end of the object with the zero mark, while depending upon visual estimation at the other end. There is little point in being much more careful at one end than at the other. Statistical theory shows that if the random error at one end is reduced to approximately one-third that of the other end, the combined error is indistinguishable from the larger error.

Often in a calibration the applied voltage is adjusted until the pointer of the instrument is judged to be over the desired scale mark, and the voltage is measured with a standard such as a potentiometer. Diagrams similar to Fig. 1 can also be drawn for this procedure, which has the advantage that there is much less uncertainty in setting the pointer over a scale mark than in estimating its position between two marks.

Systematic Errors

Even in the simplest measurement several factors can cause errors. In the calibration of a voltmeter with a potentiometer and volt box for example, neither the standard cell, potentiometer, or volt box have exactly their nominal values of emf or ratio. If the uncorrected nominal values are used, the result will be in error, on the average, by the algebraic sum of the systematic errors of the standards. Normally, for news tandards the systematic error is less than the accuracy rating. However the standards may drift with time, and may be changed in value by accidental misuse, such as by overvoltage applied to a volt box. Thus it is difficult to estimate the accuracy of a calibration when unknown systematic errors are likely to be present. It is perhaps for this reason that the 10-to-1 rule-of-thumb has come into existence, since it gives a conservative guide in such situations.

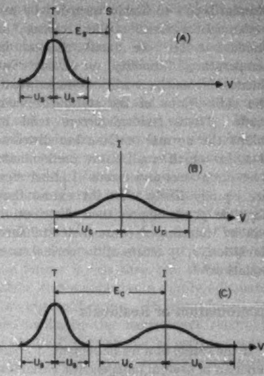


Fig. 1. Errors and uncertainties in the calibration of an instrument

In most standardizing laboratories a major uncertainty in measurements can be removed by periodically calibrating the standards that are used and by applying the corrections, illustrated as E_s in Fig. 1(A), obtained from such calibrations. This removes the determinate systematic errors from the measurements, leaving only a residue of random errors and indeterminate systematic errors. The resulting uncertainty is ordinarily so much less than the determinate systematic error that the simple precaution of obtaining and applying corrections should always be considered in a standardizing laboratory. In most cases the gain in accuracy far outweighs the time and labor involved in applying the corrections for systematic errors. If a large number of measurements are to be made with the standards, so that the time required to apply corrections is economically prohibitive, the alternatives of readjusting the standard to reduce the error or of inserting a countercorrection may be more desirable.

Residual Errors

The indeterminate residual errors in the standards that remain after corrections are applied are of several kinds. Aging and other effects cause drifts and erratic changes in the corrections to the standards. Their extent can be estimated from the differences between successive calibrations. The certainty of this estimation increases with the number of calibrations; this is why well-seasoned standards are so highly regarded by experienced workers. Other causes of residual errors are small changes in temperature, humidity, or other "influences" from the conditions under which the calibrations of the standards were made. These residual errors tend to follow the normal or Gaussian distribution, for which small errors predominate over large ones and errors of either sign are equally likely. To the extent that they do predominate, their distribution can be determined from their standard deviation, σ, by means of the normal error equation.[8,9]

Combination of Residuals

If the result of a measurement, V_r, is a function of several independent variables, so that

$$V_r = f(V_1, V_2, \ldots, V_n) \tag{1}$$

and if each variable, V_i, has a standard deviation, σ_i, which is much smaller than the mean, the standard deviation of V_r is given approximately by[3]

$$\sigma_r = \left[\sum\left(\frac{\partial V_r}{\partial V_i}\sigma_i\right)^2\right]^{1/2} \tag{2}$$

Usually the result of a measurement is the product of several factors. For example, the voltage measured with a potentiometer and volt box is $V = NDE/C$ where N is the ratio of the volt box, D the reading of potentiometer dials, E the standard cell emf, and C the setting of the standard cell dial of the potentiometer. If the exponent of each factor is $+1$ or -1, application of equation 2 leads to the much simpler expression,

$$S_r = \sqrt{\Sigma S_i^2} \tag{3}$$

where S_i is the fractional standard deviation of the factor, V_i; i.e., $S_i = \sigma_i / V_i$, and S_r is the fractional deviation of the result; i.e., $S_r = \sigma_r / V_r$. See reference 3 for the treatment of other relationships and of correlated errors.

If each factor has a normal distribution with a standard deviation much less than its mean, the result will also have a distribution very close to normal. (The result approaches a normal distribution asymptotically as the number of independent factors is increased, even if the factors do not have normal distributions.) Thus if many measurements were made under randomly varying conditions, and if determinate systematic errors were removed by applying corrections, the fractional standard deviation of the result of an individual measurement would be S_r. One would expect, from tables of the integral of the normal error curve,[9] that only three measurements in a thousand would be in error by more than $3S_r$. Therefore, the fractional uncertainty, U, of a single measurement can be taken with considerable confidence to be $U = 3S_r$.

Application to Instrument Calibration

In the calibration of a given instrument only a few readings are taken at any given scale point and many of the indeterminate residual errors in the standards are probably constant during the test. Nevertheless, in a standards laboratory the range and effect of these residuals can usually be determined fairly well from supplemental tests of the standard, and the fractional standard deviation, S_i, under the laboratory conditions, can be estimated for each. Usually the residuals are uncorrelated, i.e., independent of each other. Because of this they should be combined in quadrature as in equations 2 and 3 to obtain the residual error of the result. As indicated in Appendix I, this is much more realistic than the pessimistic method often used: directly adding the magnitudes, without regard to sign, of the extreme values of each residual.

In addition to the residual errors in the standards there are also residual errors in the instrument under test because of fluctuations in the temperature and because of other influences such as self-heating and zero shift. There are also random errors arising from the difficulty of setting the pointer of the instrument over the desired scale mark. These must also be considered in estimating the uncertainty of the result. It can be shown that ordinarily these are also residual errors in multiplicative factors, so that equation 3 applies.

The fractional uncertainty contributed by the standards is

$$U_s = 3\sqrt{\Sigma S_{si}^2} = 3S_s$$

That contributed by the instrument under test is

$$U_c = 3\sqrt{\Sigma S_{ci}^2} = 3S_c$$

where S_{si} and S_{ci} are the fractional standard deviations of individual factors affecting the standards and the instrument respectively.

As indicated in Fig. 1(C), the total uncertainty is dependent on both U_c and U_s. Since I and T are nearly equal, the total fractional variance, S_t^2, is the sum of the component fractional variances, S_c^2 and S_s^2. Therefore the total uncertainty is

$$U_t = 3S_r = 3\sqrt{S_s^2 + S_c^2} = \sqrt{U_s^2 + U_c^2} \tag{4}$$

The total uncertainty in the test arising from random errors is the range on each side of the measured value within which the true value is believed to be; i.e., the observer considers it extremely unlikely that the result is in error by more than this amount. The estimated accuracy of the test can then be defined logically as U_t.

This method of determining the accuracy is illustrated for the test of a 0.1% instrument in Appendix II and for a digital voltmeter in Appendix III.

If the known systematic error, E_s, determined from the calibration of the standard is not applied, it can be shown that a reasonable choice for the estimated accuracy of the test is

$$U_o = |E_s| + U_t \tag{5}$$

where $|E_s|$ is the magnitude of E_s without regard to sign. One would expect that the measured correction of the instrument under test would be in error by more than U_o only 0.15% of the time.

Discussion

If U_s is less than $U_c/3$, then U_t differs from U_c by less than 10% because of the quadratic relationship of equation 4. Under these conditions there would be little gain in the accuracy of the test if more accurate standards were used. (U_t differs from U_s by only 12% even if $U_s = U_c/2$. This ratio is adequate in many cases.) In contrast, from equation 5, the known systematic error, E_s, of the standard can be neglected; i.e., correction for it need not be applied, only if it is less than $U_t/10$.

If periodic calibrations of the instrument establish that the fractional standard deviation of the corrections is S_v, the uncertainty in using the instrument may be taken as $U_v = 3S_v$. Then the uncertainty of the standard, U_s, need not be less than $U_v/3$. Under these conditions, more accurate standards would not greatly increase the useful accuracy of the instrument. Some digital voltmeters which have high precision but relatively poor stability illustrate this.

These relations, which apply when systematic errors are eliminated or corrected, can be used to govern the choice of the type and accuracy of the standards for a given test, rather than the 10-to-1 rule-of-thumb. Note that the minimum accuracy ratio of 3 to 1 recommended for U_c/U_s or U_v/U_s is the ratio of uncertainties (usable accuracies), but not necessarily of rated accuracies. Often, however, when corrections for systematic errors are applied, the uncertainty of an instrument which is conservatively designed is about half the accuracy rating. To the extent that this is true for both the standard and the test instrument, the same limiting ratio of 3 to 1 would apply to their rated accuracies.

Crow has shown[10] that in a hierarchy of laboratories arranged in echelons of accuracy, the most economical ratio of accuracy between echelons for a given total error in the last echelon depends chiefly on the ratio of the number of laboratories in successive echelons and on the relationship between cost and accuracy. He concludes that when only random errors and uncorrelated systematic errors are present, the most economical ratio of accuracies between echelons frequently ranges from 2 to 4. In the notation used here this ratio is U_t/U_s. The limiting value of 3.2 corresponding to $U_s=U_c/3$ is well within this range.

Usually the fractional standard deviation of each residual can be determined only rather roughly. However, as equation 3 indicates, only the largest factors appreciably affect the result; these should be determined as accurately as possible, without much attention to the others. As a corollary, most attention should be given to reducing these if better accuracy is required in the test of an instrument. The uncertainty limits and ratios advocated here are very conservative. The less conservative limiting ratio of $U_s=U_c/2$ can often be used.

It is sometimes impractical to determine separately the magnitudes of the factors which cause uncertainties in the many types of instruments to be calibrated, and there is always some danger that a significant factor may be overlooked. An alternative procedure is to make repeated calibrations at one or two scale points of typical instruments of each accuracy rating, with different observers on different days and at different locations and under different conditions of such influences as self-heating, using the best standard that is available. For each instrument compute the fractional standard deviation of the n measured corrections as $S_t=b\sqrt{[\Sigma(C_i-C_a)^2]/(n-1)}$, where C_i is an individual correction and C_a is the average correction. The factor, b, is obtained from χ^2 tables (see, for example, Table VIII of reference 9) to obtain the upper limit of the standard deviation with a confidence coefficient of 0.90, when only a few readings are made with a given instrument. This factor is 2.37 for five readings and 1.64 for ten readings. (It should also be used in determining the standard deviations of the most significant residuals in the first method.) The uncertainty in the use of the instrument is then $3S_t$. It includes a part of the uncertainty in the standards, but this is usually very small. The total uncertainty in the standards, including the effect of changes between calibrations, is probably best determined as in Appendixes II and III.

The total uncertainty as determined by one of these two methods may in many cases be a large fraction of the accuracy rating or accuracy class of the instrument. If U_c predominates, this simply means that the instrument cannot be used to better accuracy, no matter how well it is tested. However, if its corrections determined in the test are in turn applied when the instrument is used, the measurements made with it will probably not be uncertain by much more than U_t. There is thus a real gain in the calibration of an instrument if the corrections so determined are applied. If they are not, one can only place bounds on the corrections and must downgrade the performance of the instrument accordingly. For example, if the correction of a given scale point of a 0.1% instrument is determined to be +0.08% with an uncertainty of 0.05%, the instrument may or may not be within its 0.1% guarantee. In any case, it would be difficult to make measurements with it better than 0.05%. However, if the correction of +0.08% is applied and the conditions are the same as those of the test, the result is probably accurate, at the scale marks tested, to 0.05%, a distinct gain in the use of this 0.1% instrument.

Sometimes even the alternative method of estimating uncertainties is impractical. It is usually safe to assume that the uncertainty in the use of an instrument is not more than the accuracy rating of the instrument. To calibrate the instrument under this assumption, choose standards whose total uncertainty after corrections are applied is not more than one third of this accuracy rating. The stated uncertainty for the calibration is then the accuracy rating of the instrument.

Conclusions

This paper shows that in many cases the arbitrary ratio of 10 to 1 can be reduced to about 3 to 1 if each standard is calibrated periodically, and if the corrections so determined are applied when the standard is used to calibrate other instruments. The application of such corrections largely eliminates the systematic error of the standard. The error caused by the standard in calibrating the test instrument is then only the combination of much smaller residuals, such as the effects of small variations in temperature, frequency, other influences, and uncertainties in reading the standard. These residuals or indeterminate errors can usually be estimated and a value assigned to each by an experienced worker. They are combined in quadrature with similar errors which occur in the instrument under test to determine the uncertainty or the limit of error in the calibration.

The results include the uncertainty in both the standards and in the test instrument. Ordinarily, if the uncertainty of the former is less than one third of the latter, the standards are amply adequate for the calibration. If cost functions are available the more exact relations given by Crow[10] should be used to assure that the desired accuracy of the measurements will be obtained at minimum cost in a hierarchy of laboratories.

This method of analyzing errors is illustrated by the calibration of electric instruments because the author's experience has been in this field. The principles should be applicable to the calibration of other kinds of instruments and other types of measuring apparatus as well.

Appendix I. Combination of Residual Errors

The root-mean-square or quadrature addition of uncorrelated errors is now fairly well accepted for extensive investigations in which many measurements are made under varying conditions of the causes of residual errors and their average is taken. However, it is not so readily apparent for short-time measurements, such as calibrations, in which many of the residuals are constant but unknown. Because they are constant during a calibration, it has been customary to consider them as systematic errors and to add the magnitudes of the extreme values of each, without regard to sign, to determine the maximum error that is possible in the results. However, over a long period of time, such as the time between calibrations of the standards, one would expect each of these residual errors to vary about zero in a normal manner. Each is an indeterminate error which cannot be evaluated at each use of the standard. By supplementary tests, however, the fractional standard deviation, S_i, of each can be evaluated at least roughly under the conditions in which the standard is used. One would expect that each error would exceed $3S_i$ only 0.3% of the time. Residual errors are almost always uncorrelated; that is, completely independent of each other. If such uncorrelated standard deviations are combined in quadrature, as in equation 3 to determine S_r, one would also expect that over a long period of time the resultant error would exceed $3S_r$ only 0.3% of the time.

On the basis of direct addition of errors, the limit of error of an individual factor would probably be estimated as $3S_i$, and the total for all factors would be $U_a = 3\Sigma S_i$. If now, as a simple example, five factors with equal residuals are considered, then $U_a = 15S_i = 6.7S_r$, since $S_r = \sqrt{\Sigma S_i^2} = \sqrt{5}S_i$. From tables of normal error function one would expect that the estimated error $6.7S_r$, obtained by direct addition of the components should be exceeded only 6×10^{-11} part of the time. Of course perfectly normal error curves should not be expected, but this example indicates that this method leads to an estimate of error which can be unjustifiably high.

After all known corrections are applied, there may also be indeterminate systematic errors which remain constant even over long periods of time. They are difficult to evaluate except by comparisons with other standards, preferably of a different kind. Normally they are also uncorrelated, and estimates of their magnitudes should therefore be combined in quadrature to obtain the best estimate of the resultant error.[8,10] However, they are negligible in most calibration work in comparison with the random errors of the instrument being calibrated. Exceptions are caused chiefly by mistakes in procedure rather than by constant errors in the standards.

Appendix II. Calibration of a 0.1% "Laboratory Standard" Voltmeter

The voltmeter is calibrated with a d-c potentiometer, standard cell, and volt box which in turn have been calibrated periodically. Corrections, determined from the latest calibration, are applied to their values. The standards are frequently cross-checked with others to guard against changes from overloads and other abuse.[11]

The test is made by first setting the pointer of the instrument over the zero mark and then adjusting the voltage to deflect the pointer over the desired scale marks in succession, measuring each voltage with the potentiometer. The instrument is not allowed to warm up before the test because normally time cannot be allowed for this in the use of the instrument. The precision of setting the pointer is known from many previous repeated settings of similar instruments in the laboratory.

Self-heating and other influences are estimated from prior tests of instruments of the same type and from the limiting values given in ASA Standard *C39.1*.[12] The effects of sustained operation, self-heating and zero shift, can be surprisingly large. A limit of 0.2% is permitted in *C39.1* for multirange a-c instruments of 0.1% accuracy class at 80% of full scale. Tests have shown that changes of as much as 0.1% can be obtained in some instruments in 20 minutes of sustained operation. An estimated standard deviation of 0.02% seems reasonable.

The temperature of the particular laboratory in which the tests are made is controlled with a standard deviation, s, found from thermographs, of 0.5 C (degrees centigrade). (The thermal shielding of the cases of the standards and instrument under test undoubtedly reduces the short-time variations of temperature caused by cyclic operation of the heating- or cooling system, but not the long-time variations caused by changes in heat loads within the room.) The temperature coefficients of the many resistors in the potentiometer and volt box are normally unknown, but will rarely exceed $\pm3\times10^{-6}$ per C for manganin and other high-quality resistance alloys. We may expect that if many resistors were measured, the average temperature coefficient would be practically zero and the standard deviation, c, would be about 1×10^{-6} per C. The fractional standard deviation of the resistance of an individual resistor caused by temperature variations is then $S = cs = 0.5\times10^{-6}$. The fractional standard deviation of the temperature coefficient of the instrument under test is considered to be 1×10^{-4} per C, so that for this instrument $S = 0.5\times10^{-4}$.

It is assumed that the standards have been tested often enough so that the change to be expected between calibrations can be estimated, and that the value of the corrections from the latest tests of each are applied to eliminate the determinate systematic errors. The percentage standard deviation of the corrections of each standard is computed as $S_k = \sqrt{[\Sigma(C_{i+1} - C_i)^2]/(n-1)}$ where C_i and C_{i+1} are the fractional corrections in tests i and $i+1$, and n is the number of tests. The value of $U = 3S_k$ is a good measure of the maximum change to be expected between calibrations, and therefore the uncertainty in the use of the standard, even if a drift is superimposed on random fluctuations.

This method of calculating deviations by successive differences was used by Silsbee because of its simplicity and because it estimates the desired measure of the expected changes reasonably closely. The alternative of determining separately the drift and the random fluctuations by the method of least squares is complicated, because drifts in standards are not usually linear with time. Simon discusses the significance and use of this "mean square of the successive differences" in ordnance work.[13]

If n is small, i.e., the standard has been calibrated only a few times, S_k is itself quite uncertain and should be multiplied by the factor, b, as previously indicated. However, in such cases it is probably best to use the full uncertainty given in the calibration report or certificate for that standard, since it usually includes an allowance for possible changes in value during the recommended period between calibrations. After a number of calibrations of the standard have been made, the successive differences provide the best estimate of the changes to be expected.

The estimated value of the fractional standard deviation for each factor believed to be significant in this test, and the calculated uncertainties, are given in Table I.

It is evident from the table and computations that since $U_c = 0.081\%$ and $U_t = 0.084\%$, the standards are amply accurate for the test of this instrument. More accurate standards would not improve the accuracy appreciably. The results of the test may be considered to be accurate to 0.08%; i.e., it is extremely unlikely that they will be in error by as much as this. This may not at first seem adequate for a 0.1% instrument. However, on this rather conservative basis better accuracy is not possible. Indeed, in the use of the instrument the additional errors of estimating the position of the pointer between division lines may easily equal this.

Table I. Uncertainties in the Test of a 0.1% Voltmeter

Factors	$S\times10^4$
A. Instrument under test (at full scale):	
1. Setting pointer on zero mark (one setting)	1.4
2. Setting point on scale mark (average of two settings)	1
3. Self-heating and zero shift	2
4. Ambient temperature influence	0.5
5. Magnetic field influence	0.5
B. Standards:	
1. Changes since last calibration	
(a). Standard cell (calibrated yearly)	0.4
(b). Potentiometer (calibrated biennially)	0.3
(c). Volt box (calibrated biennially)	0.5
2. Ambient temperature influence	
(a). Standard cell	0.05
(b). Potentiometer	0.05
(c). Volt box	0.05
(d). Hysteresis effect in standard cell	0.2
3. Setting potentiometer current and dials	0.2

Calculations

$S_c = 2.7$

$S_s = 0.7$

$U_c = 3S_c = 8.1$ (0.081%)

$U_s = 3S_s = 2.3$ (0.023%)

$U_t = \sqrt{U_c^2 + U_s^2} = 8.4$ (0.084%)

In the Electrical Instruments Section of the National Bureau of Standards an instrument-reading projector greatly reduces the uncertainty in setting the pointer, and a definite warm-up procedure is specified. At least two such settings are made at each scale point. A potentiometer of special design is used and the standards are well known. Thus the results at the scale points tested are confidently believed to be accurate to 0.05%.

Appendix III. Calibration of a D-C Digital Voltmeter

Most d-c digital voltmeters of better than 0.1% accuracy are in reality automatic self-balancing potentiometers with self-contained voltage dividers or volt boxes. They have the inherent advantages of potentiometers and the convenience of automatic operation, but require extremely reliable components to make full use of these factors. Because of this they are generally checked rather frequently at one or two points, often by intercomparison with another like instrument. Actual calibrations are usually made with a manually operated potentiometer and volt box. We will consider such a calibration of a 4-digit instrument having an accuracy rating of ±(0.01% of reading +1 digit). This amounts to 0.11% at the lowest reading on each range (e.g. 1.000 volt) and 0.02% at the highest reading (e.g. 9.999 volts).

Many instruments of this type have a resolution approaching 0.01%; i.e., they do not respond to smaller changes in voltage. They contain switch-operated resistors, usually in a Kelvin-Varley circuit,

Table II. Uncertainties in the Test of a Digital Voltmeter

Factors	$S \times 10^4$
A. Digital voltmeter (at full scale):	
1. Resolution	0.3
2. Ambient temperature influence	
(a). Voltage reference	0.1
(b). Kelvin-Varley resistors	0.1
(c). Voltage divider	0.1
3. Self-heating influence	
(a). Voltage reference	0.2
(b). Kelvin-Varley resistors	0.2
(c). Voltage divider	0.2
4. Noise and pickup	0.3
B. Standards:	
1. Changes since last calibration	
(a). Standard cell	0.1
(b). Potentiometer	0.2
(c). Volt box	0.3
2. Ambient temperature influence	
(a). Standard cell	0.05
(b). Potentiometer	0.05
(c). Volt box	0.05
3. Setting potentiometer current and dials	0.1

Calculations

$S_c = 0.63$

$S_s = 0.40$

$U_c = 3S_c = 1.9$ (0.019%)

$U_s = 3S_s = 1.2$ (0.012%)

$U_t = \sqrt{U_c^2 + U_s^2} = 2.3$ (0.023%)

with an unsaturated cell or Zener diode as the voltage reference. The components are generally subject to higher self-heating than those of a manually operated potentiometer. In the following analysis it is assumed that the uncertainty, $3s$, is ±3 C, and that for these resistors $c = 2 \times 10^{-5}$ per C.

Many digital voltmeters are sensitive to erratic changes in the measured quantity and to induced voltages or "pickup." The galvanometer used with a manual potentiometer has a period of 2 to 4 seconds and tends to average out fluctuations and induced voltages, but the electronic detector of a digital voltmeter often does not. Thus a very stable d-c supply and shielded test circuits are required. An uncertainty factor for the residual noise and pickup has been included in the analysis.

It is assumed that the unsaturated cells used with the standard potentiometer are checked frequently with a bank of saturated cells maintained in the laboratory, and that they are protected against sudden temperature changes to reduce hysteresis effects. The volt box is calibrated at low and rated voltage so that its self-heating influence is known and corrected for. The volt box and potentiometer are assumed to be somewhat more stable than those used in the test of the 0.1% instrument and the potentiometer current is adjusted more frequently.

The estimated value of the fractional standard deviation for each factor, believed to be significant, and the calculated uncertainties are given in Table II. These values are given as examples only, and should not be considered to be typical of all digital voltmeters.

It is evident that even in this case, where the usable accuracy ratio is less than 2 to 1, the uncertainties in the standards do not degrade the accuracy seriously, even at the full-scale reading of the digital voltmeter where its least count, ±1 digit, is the smallest percentage of the reading.

Five-digit voltmeters with stated accuracies of ±(0.01% of reading +1 digit) (i.e. 0.02% to 0.011%) would impose much more severe requirements on the standards if their stability matched their accuracy and resolution. The largest residual affecting U_s is in the volt box. For tests of such digital voltmeters it may be necessary to construct a special voltage divider instead of using a commercial volt box. Such a divider would be constructed of resistors which could be intercompared with a direct-reading-ratio-set and could then be connected in series and parallel combinations to obtain several ratios. Insulation resistance, self-heating effects, and connection errors must of course be evaluated in determining its accuracy.

References

1. Electrical Measurements (book), F. K. Harris. John Wiley & Sons, Inc., New York, N. Y., 1952, chap. 1.

2. Theory of Measurements (book), A. Palmer. McGraw-Hill Book Company, Inc., New York, N. Y., 1912.

3. The Theory of Error (book), Y. Beers. Addison-Wesley, Reading, Mass., 1957.

4. Describing Uncertainties in Single Sample Experiments, S. J. Kline, F. A. McClintock. *Mechanical Engineering*, New York, N. Y., vol. 75, Jan. 1953, p. 3.

5. The Reliability of Measured Values, C. Eisenhart. *Photogrammetric Engineering*, Washington, D. C., vol. 18, June 1952, p. 542.

6. Basic Electrical Measurements (book), M. B. Stout. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1960, chap. 2.

7. Precise Comparison Method of Testing A-C Watthour Meters. A. W. Spinks, T. L. Zapf. *Journal of Research*, National Bureau of Standards, Washington, D. C., vol. 53, Aug. 1954, p. 95.

8. Statistics for Chemists (book), W. J. Youden. John Wiley & Sons, Inc., 1951.

9. Statistics Manual (book), E. L. Crow, F. A. Davis, M. W. Maxfield. Dover Publications, New York, N. Y., 1960.

10. An Analysis of Accumulated Error in a Hierarchy of Calibrations, E. L. Crow. *Transactions*, Professional Group on Instrumentation, Institute of Radio Engineers, New York, N. Y., vol. I-9, 1960.

11. Discussion of Precision in a Meter Checking Program, R. E. McCallum. *ISA Journal*, Pittsburgh, Pa., vol. 2, 1955, p. 469.

12. Electrical Indicating Instruments. *ASA C39.1-1959*, American Standards Association, Inc., New York, N. Y., 1959.

13. Engineer's Manual of Statistical Methods (book), L. E. Simon. John Wiley & Sons, Inc., 1941, chap. 12.

◆

Discussion

J. M. Vanderleck (Hydro-Electric Power Commission of Ontario, Toronto, Ont., Canada): The paper shows that in many cases it is sufficient for the standard instrument to be three times as accurate as the instrument which is being tested, and that it is often advisable to apply corrections when using the standard and the test instrument. More emphasis, however, should be given to the requirement for a standard calibration procedure, because the uncertainty in a calibration has little meaning without a precisely defined procedure. The procedure should specify standard operating conditions in order to limit the ranges of the variables influencing the test results. The definition of total uncertainty in a calibration should be clearly stated for the instrument user's benefit. The paper defines the total uncertainty, U_t, in the test (calibration), as the range on each side of the measured value within which the true value is believed to be. Furthermore, the paper indicates the only three measurements in 1,000 are likely to be in error by more than the total uncertainty.

Provided with the foregoing information, the user of the test instrument can adopt the standard test procedure in making a measurement, and thus estimate the uncertainty of the final result. In this instance, the uncertainty in the calibration of the instrument is U_t, and the same amount of uncertainty from the true value will be present when a measurement is made with the instrument. The true value is not known, but the correction of the instrument is given by the calibration and should be applied. Consequently the uncertainty, U_t, occurs twice in the procedure. The two uncertainties should be combined in quadrature, giving a total uncertainty of the final result equal to $\sqrt{2}\, U_t$.

Because it is impractical to improve on calibration procedures in the use of a test instrument, $\sqrt{2}\, U_t$ may be considered the best practicable accuracy. More commonly, the conditions of final use give uncertainties much larger than U_t, in which U_t may have little effect on the accuracy of the final measurement.

F. L. Hermach: We appreciate the need for standard calibration procedures. As indicated in the paper, there is no real substitute for the knowledge and experience required to eliminate mistakes. Written procedures enable such knowledge and experience to be transferred to others more readily, and can also be used to set limits on the factors on which the uncertainty depends.

If the 3-to-1 ratio advised in the paper is adopted, the total uncertainty, U_t, in the calibration is almost entirely caused by the uncertainty, U_c, in the instrument under test. If the instrument is used with the same care under the same condition shortly after the test, the uncertainty in its use at the same scale points is, as Mr. Vanderleck indicates, $\sqrt{2}U_t$, since U_c is the rms (square root of mean square) sum of uncertainties arising from setting the pointer, from self-heating and ambient temperature changes, etc., which affect both the calibration and the use of the instrument. However, as Mr. Vanderleck implies, an instrument is seldom used under conditions as closely controlled as in a standards laboratory so that the uncertainties in use are somewhat larger. In addition, after a time the additional uncertainty from drifts and changes in the instrument itself must be included, to make the total uncertainty U_t, as indicated in the paper.

A 20,000-Ampere Mechanical Switch

PETER M. NEWGARD
NONMEMBER AIEE

IN JUNE 1958, the Lawrence Radiation Laboratory contracted with Stanford Research Institute for the development of a mechanical crowbar-type switch. At that time Radiation Laboratory personnel were conducting experiments using the basic high-power *L-R-C* (inductance-resistance-capacitance) circuit shown in Fig. 1. In this circuit, the ignitron V_2 is subjected to excessive duty when heavily loaded. This paper describes a high-capacity fast-acting mechanical switch designed for this special application.

The Problem

The basic resonant circuit, shown in Fig. 1, is used to produce a pulse of direct current through the inductive load L. In operation, the capacitor bank C is charged over a period of about a minute by closing switch S_1. At some later time, after the capacitors are charged, ignitron V_1 is fired. Voltage E drops through zero to a negative value in about 5 ms (milliseconds), at which time the current through L reaches its maximum value of 20,000 amperes. When voltage E drops to −200 volts, ignitron V_2 fires to short-circuit across the inductor and its lead resistance. The inductor tries to maintain a constant current circulation of about 20,000 amperes through L, R, and ignitron V_2. This current gradually decays due to the effect of R.

Ignitron tubes have been employed in most applications for the switching operations at V_1 and V_2. They appear most feasible for moderate connection times at low voltages. The most serious limitations of ignitrons for this application are high firing-time jitter, low-voltage rating, generation of radio-frequency noise, and tendency for arc transfer to the tube walls at high current densities or long conduction periods.

A mechanical crowbar relief switch S_2 was suggested to relieve ignitron tubes in crowbar applications. The mechanical switch closes as soon as possible after the crowbar ignitron begins conduction and current transfers to it, relieving the ignitron of further duty. This eliminates arc transfer to the ignitron tube walls due to excessive conduction time.

Specifications for a mechanical switch developed by Stanford Research Institute for this application are as follows:

1. Peak open-circuit voltage, 40,000 volts.
2. Peak current, 20,000 amperes.
3. Maximum voltage drop, 200 volts.
4. Series ignitron V_1 trigger time to be controlled by switch, and adjustable from 5 to 10 ms before switch closure.
5. Maximum firing jitter to be ±5% of trigger time. (This is ±250 μsec (microseconds) at a 5-ms trigger time.)

In actual practice the trigger time is determined by firing jitter. This time is adjusted so the switch will close as soon as possible after ignitron V_2 starts conducting, while still insuring that the mechanical switch will not close before voltage has become zero. Any error that would cause the switch to close prematurely on an excessive voltage would produce severe arcing and probably destroy the switch.

Since the unlatching time of a mechanical switch is usually long, it is necessary to program the experiment to the switch. Programming is accomplished by sensing the switch position part way through its travel and using this signal to trigger the series-switching ignitron V_1. This requires that the mechanical switch operation be extremely repeatable.

To be successful, a mechanical switch for this application must make exceptionally rapid and positive contact. It soon became apparent that the usual sliding contact or the snap action mechanical switch designs had serious drawbacks when both requirements of high-current loads and fast, bounce-free operation were imposed.

A switch design based on the concept of rolling contact, as illustrated in Fig. 2, was conceived to eliminate intermittent contact during the first few μsec after initial closure. Further investigations of this concept revealed that such a switch could be designed to have a sufficiently low mechanical rise time and to carry the required current without contact damage.

Method of Design

The operating principle of this switch is such that two copper rollers act as parallel short-circuiting bars to connect two rectangular copper conductors. With the rollers retracted, these conductors are normally separated by a block of Fiberglas-laminate insulation. The assembly is operated in a 150-psi (pounds per square inch) nitrogen atmosphere to minimize arcing. A contact force in the order of 1,000 pounds is necessary to provide adequate real contact area. This force is obtained by the elastic deformation of a yoke connecting the two rollers.

Feasibility of the roller contact design is primarily determined by two factors. One factor is the resistance across a contact interface, which must be sufficiently low to permit the switch to carry its rated current without producing more heat than can be dissipated from the immediate region of contact. It is assumed that the contacts have enough mass to act as an infinite heat sink at ambient temperature. This assumption is warranted by the short total conduction time required and by the low power dissipation. An arbitrary assumption of the permissible rise in temperature of the contact interface will lead directly to a maximum power dissipation at the interface. This will in turn specify a maximum allowable voltage drop and maximum allowable contact resistance. Following this line of reasoning results in an allowable contact resistance of 5 μohms to maintain less than 100 degrees centigrade rise in temperature at the contact interface.

With electrolytic tough pitch copper used as the contact material, a force of 1,000 pounds is required to provide sufficient real contact area to maintain a sufficiently low contact conductivity. In practice, copper is the most reasonable of readily available materials because of its high electrical and thermal conductivity combined with a reasonable shear strength.

Actual contact temperatures are still unknown. Nonconducting films will increase the temperature, but movement of the rollers will reduce it. In practice the contacts were gold-plated. This increases contact area, eliminates nonconductive oxidation, and reduces apparent resistivity of the contact materials. These factors all tend to reduce the interface temperature rise.

The other factor that determines basic feasibility is the instantaneous voltage gradient between contacts prior to closure. In this particular application, the voltage across the contacts decreases from a posi-

---

Paper 61-134, recommended by the AIEE Switchgear Committee of the Power Division and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 18, 1960; made available for printing December 19, 1960.

PETER M. NEWGARD is with the Stanford Research Institute, Menlo Park, Calif.

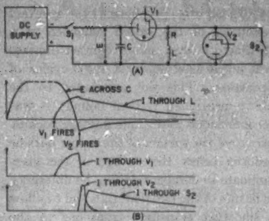


Fig. 1. A—Basic switching circuit. B—Voltage and current cycle

tive 40,000 volts to a negative 200 volts before switch closure. If switch timing is accurate, this reduction in voltage is sufficiently rapid to keep the voltage gradient within reasonable bounds until a few μsec before closure. Since the switch leads have appreciable inductive impedance, arcing across the contacts occurs only when the current through the switch is extremely low. Therefore, power dissipated in the arc is also low.

Assuming a constant roller acceleration of 11,250 inches/sec (per second) which will produce a final roller velocity of 300 inches/sec at switch contact, the voltage gradient can be calculated at any particular time prior to switch closure. Fig. 3 shows the roller approaching its contact position and defines gap d as a function of time. Since there are two series gaps per roller, there will be a maximum voltage drop of 20,000 volts/gap. Referring our time base to the time of firing the series ignitron V_1, the voltage will become zero at $t=5$ ms, and the switch will close between 1/2 and 2 ms later. At contact, there will be 100 volts across each gap as a result of the operation of the parallel ignitron V_2. This calculation will assume that V_2 fires at a time $t=5.25$ ms. This represents the maximum specified error of 5% in program time, and will produce the highest voltage gradient of any allowable program setting.

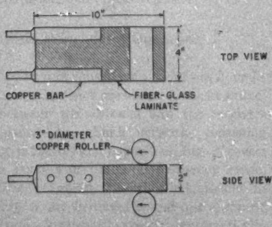


Fig. 2. Contact diagram

Since the voltage decreases toward zero approximately as a quarter cosine wave, the voltage at any time after $t=0$ and before V_2 fires is given by the expression $e = E\,(\cos 100\pi t)$ for $0<t<0.00525$ sec. The voltage gradient at any instant is approximately voltage divided by gap. Voltage gradient has been calculated for various times; see Fig. 3 for a plot of these calculated results. Any further delay in contact will produce a lower maximum gradient before the voltage zeros, therefore producing a more desirable or safer situation. From the graph it was concluded that the voltage gradient will be at a safe level until after the voltage has become zero if the contact enclosure is pressurized to 150 psi of dry nitrogen. This conclusion has been justified by laboratory test, described later.

Switch contact dimensions are confined to a rather narrow range by previous considerations of required contact area, contact force, and properties of available materials. With the contact geometry and speed of closure fixed, the switch actuator was specified. To insure repeatability of timing, it was decided that total switch travel time before contact be kept to a reasonable minimum. This would also allow a short travel which would result in a compact apparatus. A hydraulic actuator was considered, but in this case the high linear velocity would require excessively large supply lines and any oil leak near the contacts would cause serious trouble. A pneumatic actuator was chosen to provide the large accelerating force and to allow the required linear velocity. The final actuator design, shown in Fig. 4, allows the accumulation of a large amount of potential energy over a period of several seconds with an almost instantaneous release. Since the actuator requires no flow through the supply lines during its working stroke, its repeatability is nearly independent of fluid friction. Also, since the accelerating force is so large compared to friction forces throughout the system, any variation in the friction forces will have a negligible effect on repeatability.

When the switch is idle, the cocking chamber is pressurized, but both thrust and trigger chambers are open to the atmosphere through valves A and C. To operate the switch it is first necessary to cock the actuator by pressurizing both the trigger and thrust chambers. To fire the switch at any subsequent time, the trigger chamber is opened to atmosphere through valve C. As soon as trigger chamber pressure falls below a critical value, the restraining force on the trigger piston becomes less than the force on the thrust piston, and the pistons begin to move. A motion of only a few thousandths of an inch breaks the seal between trigger piston and rubber seal. This allows the pressure to equalize across the trigger piston and removes all restraining force. When desired, the roller assembly is returned to idle position by pressurizing the cocking chamber through valve B.

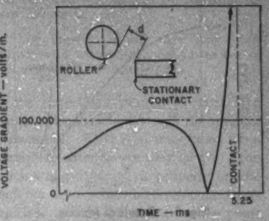


Fig. 3. Approximate voltage gradient

To insure the correct sequence of valve operation, it was necessary to build a relay and time delay control circuit. This circuit provides for either manual trigger or completely automatic cycling of the switch.

Physical dimensions of the actuator were based on the conservative assumption that no air would flow into the thrust chamber during the working stroke. It was also assumed that air in the thrust chamber expands adiabatically, a reasonable assumption in view of the short expansion time.

A hydraulic dashpot consisting of an aluminum piston in a steel cylinder was designed and constructed to absorb the kinetic energy of the contact assembly and actuator mechanism. The piston is idle during the first 4 inches of actuator stroke. It is picked up shortly after switch closure and forces oil through an orifice in the cylinder wall. The orifice is shaped in such a way that the piston acts to reduce its size by the proper amount to maintain a relatively constant deceleration over a 2-inch travel.

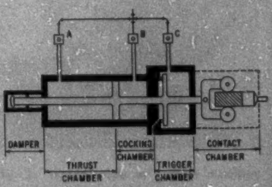


Fig. 4. Pneumatic actuator

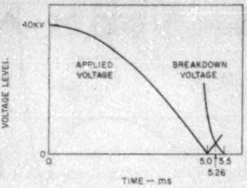


Fig. 5. Result of arc test

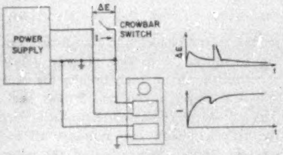


Fig. 6. Dynamic resistance test circuit and waveforms

Laboratory Testing

Mechanical Tests

With the contact chamber cover removed, the switch was operated at 300-inch/sec peak roller velocity. High-speed motion pictures of the operation show no excessive rebound from the fully extended position.

With an oscilloscope wired to observe switch contact, it was determined that the contacts do not open after the first 100 μsec of mechanical rise time.

Static Contact Resistance

Static contact resistance was measured by means of a bridge circuit with about 100 amperes passing through the switch.

With the rollers in various positions, and exposed to the atmosphere, 21 readings of contact resistance were taken over a period of 48 hours. The readings gave a range of contact resistance from 4.18 to 3.14 μohms, which is within the allowable limit.

Static contact resistance was again measured after the first 100 cycles of operation, when the readings varied from 5.30 to 2.70 μohms. These values show more variation, but are still acceptable.

Voltage Breakdown Tests

To test for breakdown with the switch open (rollers completely retracted) and the contact chamber pressurized to 150 psi, a high-voltage supply was connected across the switch contact bars. A center-tapped bleeder resistor was used to keep the switch case at ground potential and midway between the two contact bars. Voltage was increased from zero to 50 kv with no surge of current to indicate arcing.

A high-voltage supply circuit was used to determine the time at which arcing would occur during switch operation. The oscilloscope was triggered by a photocell within the contact chamber. When voltage breakdown occurred, the scope displayed an abrupt change in d-c level. Fig. 5 shows the resulting breakdown time as a function of voltage level, and also the voltage that will be applied during switch operation. Since the voltage required to produce arcing is always much greater than that to be applied during operation, it may be concluded that the switch will arc only for a very few μsec before closure.

Dynamic Contact Resistance Test

Contact resistance was measured while the switch was in operation, with the use of the circuit shown in Fig. 6. This circuit is capable of supplying about 1,000 amperes for a few milliseconds. The scope trace shown exhibits a discontinuous current and voltage. This is a result of rollers' jumping vertically away from bars after initial contact. The time of this discontinuity correlates with the natural frequency of the rollers and spring yoke.

The switch was tested at Lawrence Radiation Laboratory under a load of 18,000 amperes with 200 volts applied at switch closure. Although the contact resistance of the switch remained at a satisfactory level, minor contact damage was noted after the first 1,000 cycles of operation. This was attributed to the above-mentioned vertical bound; it was recommended that if this damage became excessive, the surface of the copper bars be reduced below the surface of the glass laminate to compensate for the difference in flexure of the two materials and to allow a smoother transition from one to the other.

During the 18,000-ampere test, the small vertical bounce resulted in some damage when the rollers actually broke contact with the stationary bars. This bounce continued for a second cycle, but with such a reduced amplitude that the contact force was only partially relieved. During this period of reduced contact force, there was a radical increase in contact resistance, but only a slight decrease in current, so that power dissipated at the contacts was also high. It was estimated that the power dissipation during this interval was equivalent to that which would result if the switch were conducting over 10^6 amperes at its normal contact resistance. This power dissipation caused no damage to the contacts, indicating that the roller contact design could be developed to conduct much higher currents while the rollers are in motion.

Conclusions

The roller contact design has been proved capable of conducting at least 18,000 amperes. A contact resistance of about 5 μohms was achieved which resulted in a voltage drop of 0.1 volt and a power dissipation of 2,000 watts during the conduction period.

The required mechanical repeatability was achieved by the pneumatic actuator.

Some of the laboratory tests give a strong indication that this design may be developed to conduct currents of the order of 10^6 amperes while the rollers are in motion.

Antenna Matching Unit for High-Frequency Vehicular Whip

J. R. GRUBER
ASSOCIATE MEMBER AIEE

G. J. SEWARD
NONMEMBER AIEE

THIS paper describes the design details and performance of an automatic impedance matching unit designed to operate with a vehicular communication equipment consisting of a receiver-transmitter and a 10-foot whip antenna.

Purpose of Equipment

The antenna matching unit, Fig. 1, was developed as a component of a military communications set covering the continuous frequency range of 30 to 75.95 mc (megacycles). The communications set, the *AN/VRC-12*, provides 920 separate channels with 50-kc spacing and operates from a 24-volt d-c vehicular battery supply. The system is designed for installation on vehicles such as jeeps, trucks, and armored vehicles.

The purpose of the antenna matching unit is to match a Signal Corps whip antenna to a 50-ohm characteristic impedance with a VSWR (voltage standing wave ratio) of less than 3 to 1 over the operating range. This insures a radiation efficiency for transmission and reception of more than 75%, whereas, without the matching unit, the efficiency is as low as 20%.

The development of the matching unit is a significant step forward in the field of broadband vehicular communications since it operates automatically as the receiver transmitter is tuned in frequency and eliminates the manual adjustments of predecessor equipments. Previous equipments contained two interdependent manual controls which had to be adjusted by the operator each time the frequency of operation was changed. This was time-consuming and difficult to do under battle conditions and in bad weather.

Basic Principle of Operation

The unit consists of a set of ten *L*- or *T*-networks which are inserted, one at a time, between the antenna and the receiver-transmitter. Each network, consisting of inductors and capacitors, Fig. 2, covers a portion of the total frequency range. The networks, shown schematically in Fig. 3, are switched in and out of the r-f (radio-frequency) circuit automatically as the communication equipment is tuned in frequency. The control voltages which accomplish this are derived from the receiver-transmitter via the interconnecting power cable and are applied to the rotary solenoid which drives switch *S551*. The antenna is thus matched for both transmitting and receiving, since both the transmitter and receiver have a 50-ohm impedance. Band change relays are also operated automatically from the receiver-transmitter. Thus, the matching unit is controlled remotely and can be installed on the vehicle at any desired location.

General Characteristics

Physical Characteristics

The physical characteristics of the antenna matching unit include the following:

1. The antenna matching unit is housed in an aluminum alloy casting which gives protection against impact.
2. The unit is provided with a watertight gasket to seal the unit for protection against weather and complete submersion. No leaks developed when the unit was submerged under water for over 2 hours.
3. The components are mounted rigidly so as to meet Signal Corps vibration requirements for vehicular equipment.
4. The entire unit operates satisfactorily over the required temperature range of −40 C (degrees centigrade) to +65 C.

Electrical Characteristics

The electrical characteristics of the antenna matching unit include the following:

1. Ten matching *L* or *T* networks are provided so as to keep the standing wave ratio below 3 to 1, referred to 50 ohms, over the frequency range of 30 to 76 mc.
2. The matching networks are designed to have an insertion loss of less than 0.1 db (decibel) and will handle r-f power of at least 100 watts.
3. The impedance will vary only very slightly with temperature due to the use of temperature-stable components.

The impedance matching takes into account the vehicle on which the antenna is mounted. That is, a unit tuned for a particular antenna position on a 3/4-ton truck will neither work properly on the jeep nor on a point on the 3/4-ton truck which is greatly different from the point where matching was accomplished.

The reason for this vehicular dependence is that the antenna current distribution and impedance depend upon the nature of the surroundings. The vehicle on which the whip is mounted actually becomes a part of the antenna circuit and represents a sort of counter poise for the whip. Therefore the antenna impedance will vary for different vehicles and mounting positions. If the mounting position is changed the matching networks have to be retuned for optimum performance.

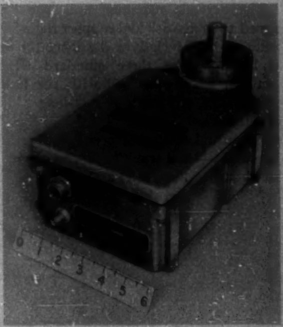


Fig. 1. Antenna matching unit for radio set AN/VRC-12

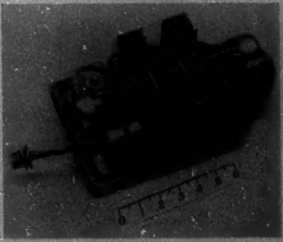


Fig. 2. Internal view of antenna matching unit

Paper 61-47, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1960. Manuscript submitted June 14, 1960; made available for printing December 2, 1960.

J. R. Gruber and G. J. Seward are with the Avco Corporation, Cincinnati, Ohio.

The work described herein was performed under Signal Corps Contract *DA-36-039-SC-72794*.

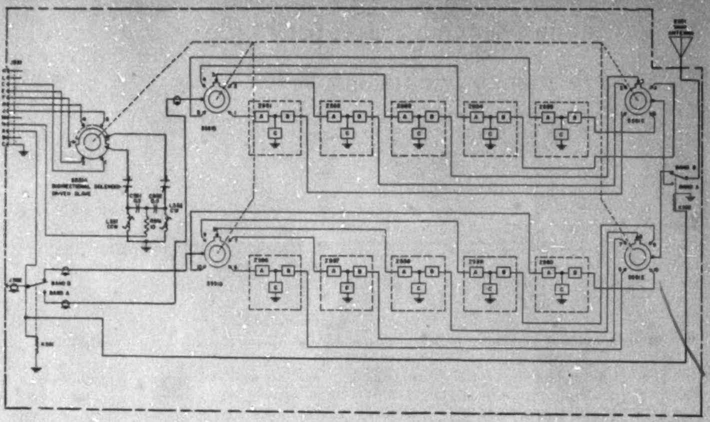


Fig. 3. Schematic diagram of antenna matching unit

Design Details

MOUNTING

The antenna matching unit was designed to mate with and support the whip antenna base including the insulator as in Fig. 4. The three standard whip sections, when combined with the base unit of Fig. 4, form a 10-foot-long whip antenna.

The usual mounting positions for the jeep and 3/4-ton truck (weapons carrier) are shown in Figs. 5(A) and (B).

A somewhat different arrangement is necessary in a tank installation where the matching unit is placed just inside the armored shell and the antenna is inserted from the outside. No adapters or special fittings are necessary for the tank installation.

RADIO-FREQUENCY BAND SWITCHING

Switching between the high and low bands is accomplished by means of a coaxial switch, *K551*, on the receiver-transmitter side and a miniature relay on the antenna side of the rotary solenoid switching unit. By use of the 50-ohm

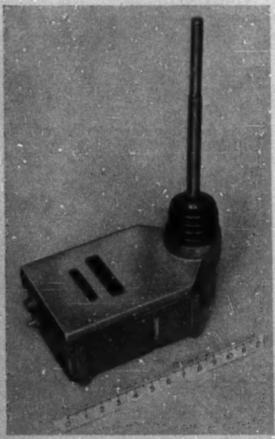


Fig. 4. Antenna matching unit with mast base AB-15/ GR

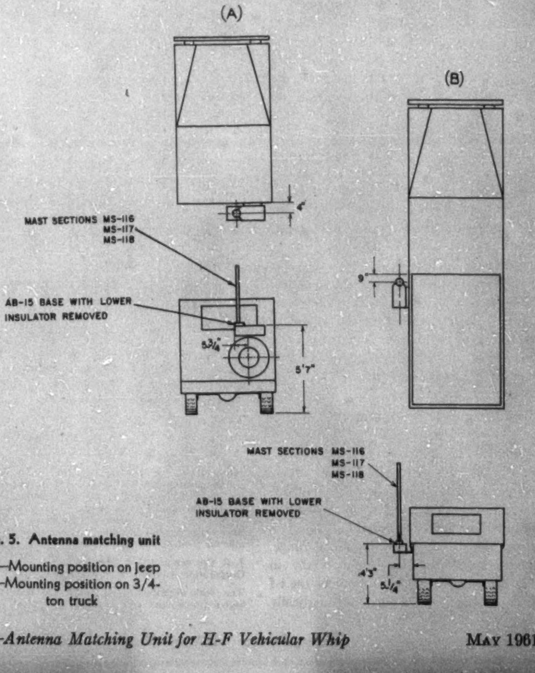
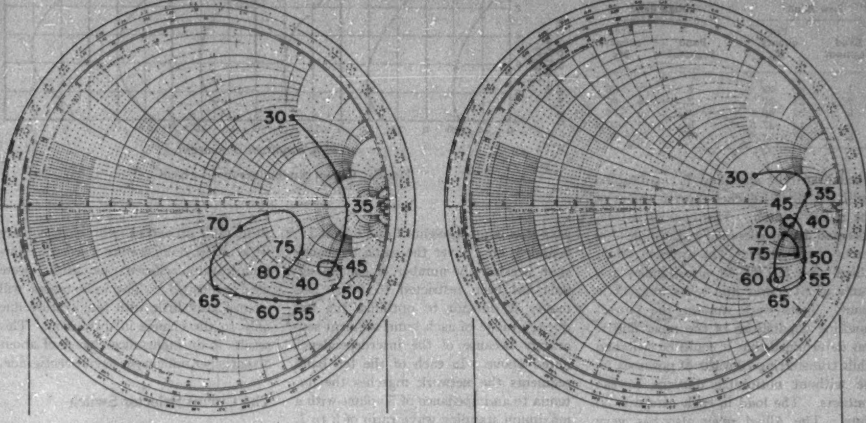

(A)
(B)
MAST SECTIONS MS-116
MS-117
MS-118
AB-15 BASE WITH LOWER INSULATOR REMOVED
4"
5 3/4"
5'7"
9"
MAST SECTIONS MS-116
MS-117
MS-118
AB-15 BASE WITH LOWER INSULATOR REMOVED
4'3"
5 1/4"

Fig. 5. Antenna matching unit

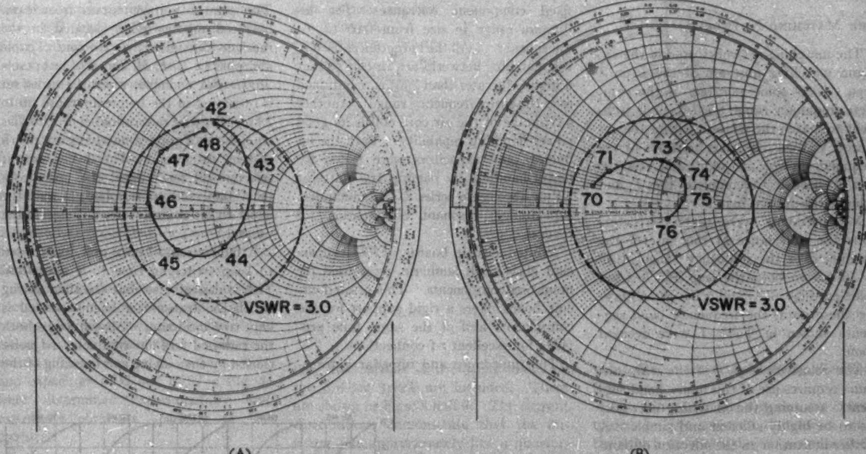
A—Mounting position on Jeep
B—Mounting position on 3/4-ton truck



(A) (B)

Fig. 6. Impedance of 10-foot whip antenna (normalized to 50-ohm impedance)

A—Mounted on left side of 3/4-ton truck
B—Mounted on jeep



(A) (B)

Fig. 7. Admittance of base-fed antenna (normalized to 20 millimhos)

A—Mounted on jeep
B—Mounted on 3/4-ton truck

Table I. Frequency Band Segments

Low Band		High Band	
Band Segment	Frequency, Mc	Band Segment	Frequency, Mc
1	30 -33	6	53 -56
2	33 -37	7	56 -60
3	37 -42	8	60 -65
4	42 -47.5	9	65 -70.5
5	47.5-53	10	70.5-76

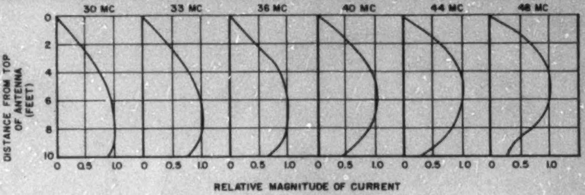


Fig. 8. Current distributions for low band

coaxial switch, shielding is maintained until the signal reaches the r-f wafers of the rotary switch. This coaxial switch, operated by a solenoid, will handle 100 watts of r-f power. The antenna relay, Allied type *KHYX-48*, also handles 100 watts of r-f power and takes very little space. Tests were run on the power-handling capabilities of the relay and it was determined that it could be switched while transferring 70 watts of power at 50 mc without noticeable damage to the contacts. The load in each case was 50 ohms. The Allied relay also has very low capacity to ground and intercontact capacitance, i.e., on the order of a picofarad. If both relays are energized, the low band networks are in the circuit. If de-energized the high band is in the circuit. Both of the solenoid-operated switches and the relay are designed for operation on a d-c supply voltage of 22 to 30 volts.

The Matching Networks

The matching networks used in the antenna matching unit are of the *L* or *T* type. Over some portions of the complete frequency range the *L* type is sufficient to match the whip to within the required VSWR. In other portions of the frequency range an additional reactive arm is needed to complete the *T* and achieve the desired match.

Measurements were taken to determine the impedance at the base of the antenna over the frequency range of 30 to 76 mc on both the jeep and 3/4-ton vehicles; Figs. 6(A) and 6(B). By observing these data, the frequency band segments shown in Table I were decided upon.

The relatively high *Q* of the whip antenna requires use of this many band segments, assuming the matching networks are to be highly efficient and simple. A further limitation on the selection of band segments is due to the type of control voltages available from the receiver-transmitter of the *AN/VRC-12*. The control voltages are all obtained from the 24-volt d-c supply and are transmitted to the matching unit via five separate conductors of the control cable. It is necessary that band segments 1 and 6, 2 and 7, 3 and 8, etc., cover the same frequency span because the number of control signals has been restricted to five. This makes it difficult to optimize the frequency range of each band segment separately because of the interdependence noted above. In each of the ten band segments the network matches the antenna to an impedance of 50 ohms with a maximum standing wave ratio of 3 to 1, using only low-loss components. Typical results are given in Figs. 7(A) and 7(B).

The fixed capacitors used in the networks are ceramic type with a *Q* generally in excess of 1,000. These capacitors range in size from 15 to 75 picofarads. Variable piston capacitors, 0–18 picofarads, are used for optimizing the match of each network and to compensate for fixed component tolerances. The inductors range in size from 0.05 to 0.4 microhenry. All the inductances used in the matching networks are air core so that a high *Q* (greater than 200) is maintained over the entire frequency range. Another advantage of the air core coil is its good power-handling capability. Both the inductors and capacitors were tested for power-handling by passing 100 watts at 50 mc through a series *L-C* (inductance-capacitance) resonant combination for several hours.

Printed circuit boards, *FF91* copper-clad glass cloth laminate, were used for mounting components. This printed circuit board offers a rigid support for the components and at the same time provides for excellent r-f connections.

The impedance and impedance transformation properties will change very little with temperature because low-temperature-coefficient capacitors were used exclusively. The inductances will change very little with temperature due only to the change in dimension. This results in an inductance change of about 20 parts per million per degree centigrade.

The Circuit Selector Switch

Remote switching of the networks in the antenna matching unit is accomplished by the use of the Ledex-driven circuit selector. This wafer-type switch rotates in discrete steps by means of a rotary solenoid used with a ratchet and detent system. The bidirectional homing action of the switch is achieved by having two sets of control wafers (see Fig. 3). The control voltage comes from cam-operated microswitches, located in the receiver-transmitter, via the control cable and pins *D*, *E*, *F*, *H*, and *J* of receptacle *J551*. For example, when the radio set is operating in the frequency range 30 to 33 mc, the wire to terminal *D* of *J551* has a potential of approximately 24 volts with respect to ground. The Ledex circuit selector will rotate until the wafer contact connected to terminal *D* and the wiper on the control wafer no longer make contact.

A bidirectional Ledex unit was chosen to keep switching time to a minimum. This is desirable so that the time during which the transmitter is mismatched is kept to a minimum. Also, it keeps both the radiated and line-conducted r-f noise caused by the making and breaking of the

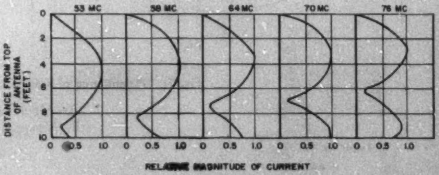
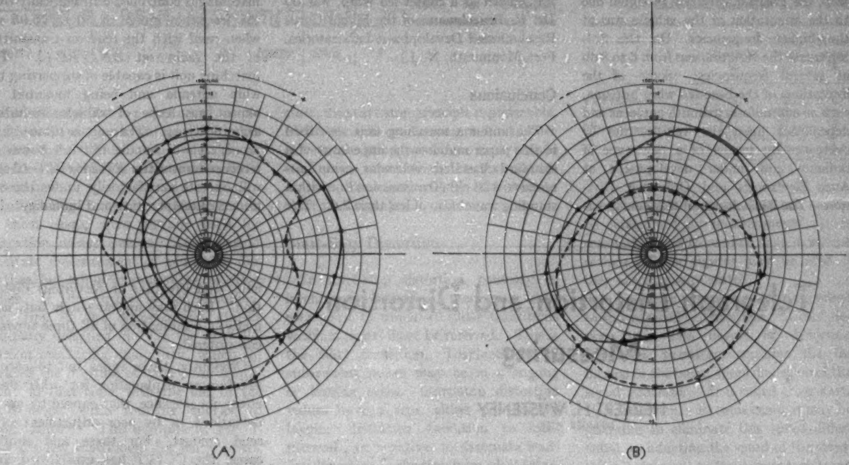


Fig. 9. Current distributions for high band



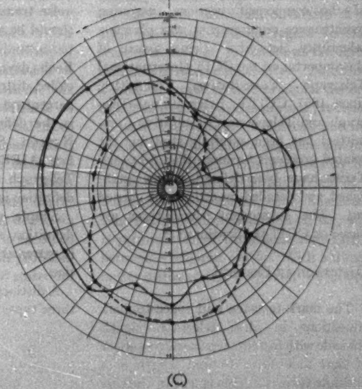


Fig. 10. Antenna pattern

O——O Data taken with antenna mounted on 3/4-ton truck

O- - - -O Data taken with antenna mounted on jeep

A—Frequency, 30 mc

B—Frequency, 53 mc (7 antennas)

C—Frequency, 76 mc

contacts controlling the solenoids to a minimum. The shorter switching time of the bidirectional Ledex is accomplished by proper use of the slave control wafer. As seen by the circuit diagram in Fig. 3, the "motor," turning the Ledex in the direction which minimizes the switching time, will always be energized. The above advantages of the bidirectional Ledex are obtained at some increase in size because space is needed for the second rotary solenoid.

The tabs used on the r-f wafers are of the short-circuit type. That is, the transmitter never "sees" an open circuit, but, for a very brief time during the switching operation, two of the matching networks will be in parallel. This time is so short, in order that the power amplifier tube of the transmitter will not be damaged.

A spark suppression network is connected across the control contacts of the rotary solenoids to increase the life of the contacts and to reduce the r-f noise generated. The components, resistor *R551* and capacitors *C551* and *C552* of Fig. 3, were determined experimentally and essentially eliminate sparking of the contacts.

Performance of the 10-Foot Base-Fed Antenna

Patterns and current distribution measurements were taken on both the jeep and the 3/4-ton vehicles. The current distributions were very similar; therefore only the current distributions taken on the 3/4-ton truck are included. These are shown in Figs. 8 and 9. The current distributions demonstrate that the antenna acts approximately like a quarter-wave radiator at the lower frequencies and a half-wave device at about 50 mc. From a field-strength standpoint the optimum frequency is about 58 mc where the current distribution represents that of a radiator 5/8 wavelength long. At higher frequencies the field strength along the ground is 1 to 3 db below the optimum value due to the actual current distribution.

The currents in the antenna are determined by the antenna dimensions and the vehicle on which the whip is mounted; they do not depend upon the matching network.

Patterns were taken on the jeep and 3/4-ton vehicle; see Figs. 10(A) through 10(C) for representative measured patterns. These patterns are dependent on both the frequency and the vehicle on which the antenna is mounted. On the jeep, the greatest variation in signal due to the orientation of the vehicle was at the higher frequencies. On the 3/4-ton truck, the variation was from 6 to 8 db at several frequencies. (Some of the limitations of the base-fed whip antenna, such as nonuniform azimuth patterns and dependence upon the vehicle, can be overcome by means of a new type of radiator now under development at Avco Electronics and Ordnance Division. The basic design of this new radiator, known as a center-fed whip, was by Dr. H. Brueckmann of the Signal Corps Research and Development Laboratories, Fort Monmouth, N. J.)

Conclusions

The antenna matching unit described in this paper matches the impedance of a standard base-fed vehicular whip antenna to a 50-ohm transmission line with a standing wave ratio of less than 3.0. The matching is controlled automatically over the frequency range of 30 to 75.95 mc when used with the receiver-transmitter of the radio set *AN/VRC-12*. The matching unit is capable of supporting the whip antenna and being mounted on almost any type of vehicle, including armored types. The unit is submersion-proof; it will handle the full power of receiver-transmitter *RT-246/VRC-12* and will operate satisfactorily under the extremes of temperature and humidity.

Telegraph Distortion and Distortion Measuring

HERBERT H. WÜSTENEY
MEMBER AIEE

THE DEFINITION of distortion with the resulting conclusions is of fundamental importance in determining the reliability of telegraph transmission circuits, judging the quality of the component sections, assessing maintenance requirements, and planning teleprinter networks. Being aware of this importance, the CCIT (International Telegraph Consultative Committee, since 1957 CCITT, International Telegraph and Telephone Consultative Committee) has been carefully studying the related problems for more than 30 years. Terms have been defined, values established, and detailed investigations carried out. The subsequent elucidations are based on CCIT publications.[1–3]

Distortion in General

The mark-to-space and space-to-mark transitions in ideal telegraph signals coincide with instants separated by periods of time amounting to unit length or a multiple thereof. This is true in the case of isochronous systems. The term isochronous, which means having the same speed of revolution and the same phase position, was introduced recently by the CCITT because isochronous is more to the point than the previously used term, synchronous, which is required for another purpose.

In start-stop telegraph systems, the start pulse transition serves as a reference point, from which the remaining pulse transitions of the respective signal should be spaced at intervals amounting to unit pulse length or a multiple thereof. Each deviation from these instants is called distortion, which is expressed as a percentage of a unit pulse length. This is the definition originally furnished by Nyquist, Shanck, and Cory[4] in their first AIEE paper on "Measurement of Telegraph Distortion" in 1927. This definition is still valid.

Simple and straightforward as this definition may be, telegraph distortion is extremely intricate to handle in practice.

A distinction is generally made between three types of distortion: bias, characteristic, and irregular or fortuitous distortion. Each may appear singly but, normally, a combination of all three types is encountered.

Obviously, these types fail to give a satisfactory explanation for other varieties; for instance, that caused by speed deviations, or by poor adjustment of a send contact. For these and similar cases, the CCITT has established and defined the term, cyclic distortion. This term is not used in practice, however, because these varieties must be treated as fortuitous distortions. This is permissible because, although its value remains constant in a single equipment, cyclic distortion varies in magnitude and sense from machine to machine.

Conventional voltage or current measurements are clear-cut. All that is required is a fundamental definition and measuring equipment designed in accordance therewith. Measuring telegraph distortion is something else, because in practical service over large distances, the amount of distortion is not the same for all pulse transitions in a series of signals. Distortion values are distributed over a smaller or larger range and peak values are infrequent. Uncertainty results as to whether peak values are included in measurements extending over a

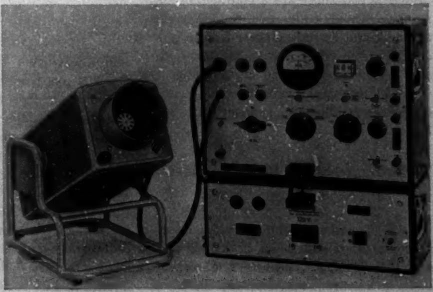


Fig. 1. Isochronous distortion meter

Paper 61-60, recommended by the AIEE Telegraph Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted September 29, 1960; made available for printing November 30, 1960.

HERBERT H. WÜSTENEY is with Siemens and Halske, A. G., Munich, Germany.

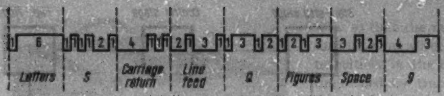


Fig. 2. Standard test code for distortion measurements and line tests

comparatively short period, and as to whether to use peak values for further conclusions. Which value should be regarded as measured value remains to be determined.

Integrating methods, bridge circuits in most cases, were initially used for distortion measurements. They permit accurate measurement of pure bias distortion, but are of limited value for other types. If the author is not mistaken, Mr. Jipp of the Siemens and Halske company designed in 1928 the first distortion measuring equipment, which indicated the position of each pulse transition; at that time, with the aid of the stroboscope with rotating sparks or neon lamps.[5,6] This equipment, Fig. 1, was the key instrument for many years, and even today it is used on a large scale, because it has gone through many further developments.

Integrating, as well as stroboscopic, methods are generally applied in isochronous systems. The amount of distortion is here identical to the range within which the pulse transitions are located, when compared with a time base corresponding to the signal transmission rate. Since the nominal instants of the transitions are not defined, only the range within which the transitions stray can be determined. Therefore, neither positive nor negative values are found in isochronous systems; merely distortion.

Isochronous measuring methods are used normally for checking and line-up of individual line sections, such as the v-f (voice-frequency) telegraph transmission channels. For checking complete circuits, the start-stop principle is preferable. Measurements are then carried out under actual operating conditions and permit continuous supervision. Further considerations, therefore, will be restricted to start-stop signals.

Start-Stop Distortion

In start-stop distortion, contrary to isochronous distortion, there is a definite point to which all pulse transitions within a signal must be referred. This is the start transition. Displacement of subsequent pulses may be in a lagging or leading sense. Start-stop distortion values have a sign, either positive for lagging distortion (addition to ideal interval), or negative, to designate leading distortion (reduction from ideal interval).[7] In the CCITT definitions and in practical operation, no difference is made between transitions from mark-to-space and space-to-mark. The term, end distortion, is therefore not contained in these definitions.

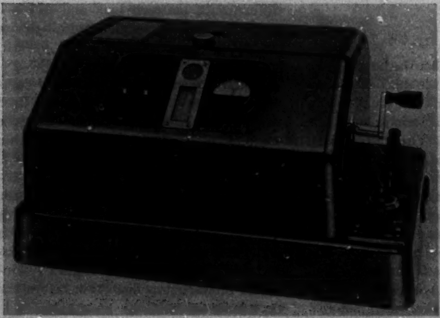
The degree of start-stop distortion of a series of teleprinter signals is, in principle, the maximum value of the distortion of all the individual pulse transitions. In practice, however, it is not advisable to include in any calculation the actual maximum value which is so infrequent that it does not markedly affect the error frequency. An understanding has been reached to regard as conventional degree of distortion that value which, within an extended period of operation, is exceeded by a certain though very small probability. The figure of 1 to 100,000 stated for this probability is so small it can hardly be determined by measurement.[8] It is reasonably safe to assume that it will be revised. Other figures may be used if they are expressly stated.

Table I. Limit Values for Isochronous and Start-Stop Distortion

		Isochronous, %	Start-Stop, %	Measured, %
1	v-f channel	10	8	6
2	v-f channels in series	18	13	11
3		24	17	15
4		28	21	19
5			25	21

A distortion measuring set is supposed to run normally at the nominal modulation rate. The degree of distortion registered, therefore, contains the influence of the mean speed deviation of the signal generator. It is called gross start-stop distortion. In some cases, it may be desirable to eliminate this speed differential by adapting the speed of the measuring instrument to the momentary mean modulation rate of the signals. The result of this measurement is the degree of synchronous start-stop distortion.

As previously stated, the figure indicating the start-stop distortion reflects the maximum individual distortion of a pulse transition without regard to the sense of distortion; it also shows the deterioration in receive reliability. To afford a more accurate picture of signal distortion it is possible, and is normal practice, to state two values: a positive one for lagging distortion and a negative one for leading distortion. This practice may be adopted for verifying the adjustment of a teleprinter transmitter, for instance. Where a start-stop distortion of, say, 4% is permissible, all transitions may range between +4% and −4%. With a transmitter adjusted for +3% and −1%, one



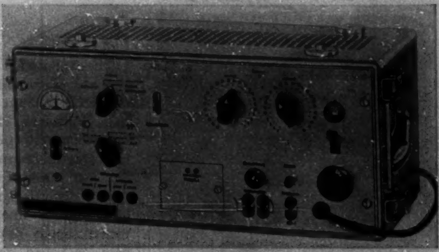


Fig. 3 (left). Electromechanical start-stop distortion monitor

Fig. 4 (above). Electronic start-stop distortion monitor

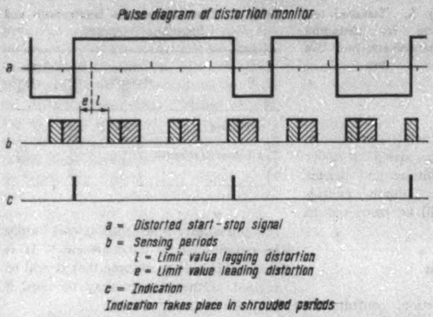

Pulse diagram of distortion monitor
a
b
c
e
l
a = Distorted start-stop signal
b = Sensing periods
l = Limit value lagging distortion
e = Limit value leading distortion
c = Indication
Indication takes place in shrouded periods

Fig. 5. Operating principle of electronic start-stop distortion monitor

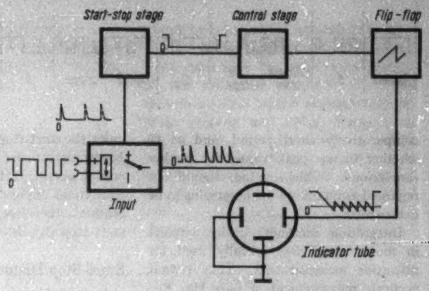


Fig. 7. Operating principle of start-stop distortion meter

might simply say that it has a distortion of 3%. The second figure, however, tells us that this adjustment is much more accurate on one of the two sides. Moreover, the signs which indicate the sense of distortion facilitate location and elimination of trouble sources.

If the signals of a teleprinter transmitter have an actual start-stop distortion of ±x%, the same signals registered by an isochronous equipment will come up to 2x%. Of course, this can be measured only with accurate speed adjustment, continuous transmission, and a unit stop pulse length. On the other hand, a start-stop distortion of ±x% may appear, at the end of a line, if only a fortuitous isochronous distortion of x% is present with the start pulse transition straying within the same range as the remaining pulse transitions. The ratio of start-stop distortion and isochronous distortion may thus vary between 1 to 2 and 1 to 1. Comparative measurements on physical circuits resulted in a value of approximately 0.7.

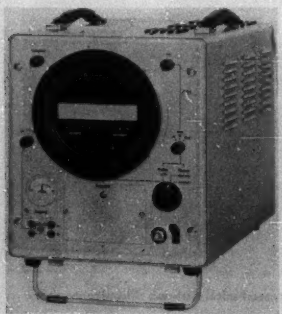


Fig. 6. Start-stop distortion meter

To obtain results permitting a comparison, a standard test code has been set up for measuring the distortion in telegraph lines and this test code seen in Fig. 2 uses these characters:

LTRS, S, CR, LF, Q, FIGS, SPACE, 9

The signals must have unit stop pulse lengths to permit measurements with isochronous as well as start-stop equipment. They have been chosen for the purpose of obtaining a highly irregular modulation pattern so as to generate an amount of distortion closely resembling that produced with normal message text.

Supervisory measurements on lines should not last too long. This, however, increases the likelihood that the actual distortion is greater than the measured value. The CCITT recommends a tentative measuring period of 30 seconds and the use of the standard test code, measuring the lagging and leading distortion, each for a 15-second period. If measured values are $+x$% and $-y$%, the approximation to the total value reads,

$$d = x + 1/3y,\ \text{if}\ x > y$$

or

$$d = y + 1/3x,\ \text{if}\ y > x$$[9]

The explanation is that the larger of the two values contains the bias distortion to the full amount while the smaller contains only the characteristic and the fortuitous distortion. The latter ones, however, are supposed to be read with only two-thirds of their peak values.

The distortion measured at the end of a line or line section, with undistorted signals impressed at the beginning, is called degree of inherent distortion. The total distortion at the end of a multisection line is smaller than the algebraic sum of the individual distortion values. The CCITT gives the figures shown in Table I for the permissible limit values of isochronous distortion and start-stop distortion.[10] Added to these, are the peak values of start-stop distortion which were measured on a large number of v-f telegraph channels without previous lineup.[11]

It should be remembered that all CCITT recommendations apply only to international circuits.

An equation recommended for distortion calculations is[12]

$$\delta\ \text{inherent} = \sum_{1}^{n} \delta_c \sqrt{\sum_{1}^{n} (\delta_b)^2 + \sum_{1}^{n} (\delta_t)^2}$$

where

n = number of sections (>1)

δ_c = characteristic start-stop distortion of a section

δ_b = bias distortion of a section measured with 1 to 1 or 2 to 2 reversals

δ_t = fortuitous distortion of a section

Measuring Equipment for Start-Stop Distortion

Start-stop distortion measuring sets manufactured in Europe are adapted to CCITT terms. Sets indicate leading and lagging distortion, but do not differentiate between bias and end distortion. The electromechanical device shown in Fig. 3 was proposed by Mr. Jipp and was the first type used.[13]

Signals are sampled by means of cams and contacts. A neon lamp is fired when a pulse transition coincides with the sampling time, which can be varied within a range from practically 0% to 100%. When adjusted for 100%, the slightest distortion of a pulse transition will fire the neon lamp. On the scale, this point is marked with "0% distortion." Each per cent displacement effecting a shortening of the sampling time permits a 1% increase in start-stop distortion before the neon lamp lights.

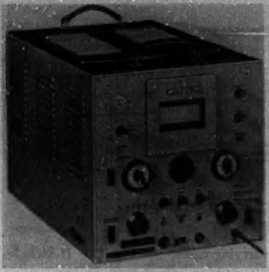


Fig. 8. High-precision distortion meter

This electromechanical equipment has been outmoded by a new electronic unit. Siemens and Halske are now manufacturing an electronic telegraph distortion monitor, seen in Fig. 4. The limit values whose exceedance is indicated by the firing of the neon lamp can be adjusted in steps of 2% between 2% and 48% by means of two rotary switches.[14]

For purposes of a statistical analysis, four outlets are provided which permit registering of these criteria by means of suitable pulse counters:

1. Number of all transitions.
2. Number of transitions exceeding the limits.
3. Number of all teleprinter signals.
4. Number of those teleprinter signals where the limit value was exceeded at least once.

Fig. 5 illustrates the operating principle. The time base is provided by a 20-kc crystal oscillator, which starts the measuring cycle with the first pulse following the start pulse transition; i.e., with a delay amounting to a maximum of 0.25% at a transmission rate of 50 bauds.

The limit values to be adjusted with the control knobs are marked as "lagging" and "leading." The ranges within which pulse transitions are indicated have been shaded. In the illustrated example, the transition at the end of the start pulse comes within the shaded portion and is therefore indicated, as well as the transition at the beginning of the third code element. This is not the case with the transitions at the beginning of the fourth and fifth pulses. The beginning of the stop pulse, however, is again located within this range and is indicated. This equipment is used chiefly for circuit supervision.

For measuring purposes, the start-stop distortion meter, Fig. 6, is available. The positions of the pulse transitions are displayed by means of a cathode-ray tube so that the magnitude of distortion can be read directly on a scale. The transitions in the two directions are distinguished by traces of different height. The scale has a length of 4 inches for ±50% distortion; i.e., 40/1,000 inch per per cent. This insures a reading error below 1%.

Controlled by a saw-toothed voltage applied to one plate pair of the oscilloscope, the beam moves from −50% to +50%, illustrated in Fig. 7, within a period equaling one signal element. This beam deflection behind the scale remains invisible. Each pulse transition generates a pulse at the other plate pair, resulting in a vertical deflection of the beam which now appears above the scale at this point.

Up to a range of 30% distortion, the measuring error will not exceed 1%. This value is only slightly exceeded at a higher distortion.

For accurate distortion measurements in the laboratory; e.g., for determining the outgoing signal distortion of a teleprinter, a special high-precision distortion meter, Fig. 8, has been developed. Distortion values from −50% to +50% or, after flipping a switch, from −10% to +10%, are displayed on the screen of a cathode-ray tube through a range amounting to 4 inches. With the switch in the latter position, 1% distortion equals a displacement by 0.2 inch on the scale.

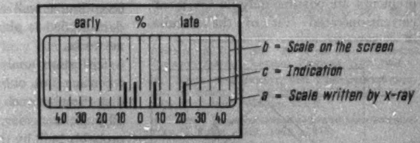
The reading is fully independent of inaccuracies in beam deflection. This has been achieved by an arrangement whereby the beam itself writes the scale under the control of pulses supplied by the crystal oscillator. This electronic scale may then be brought into alignment with the illuminated scale. The scale marking pulses appear on the screen in the form of very thin traces while the indicating pulses produce a thicker trace; see top of Fig. 9.

In one position of the switch, the beam is raised with each pulse into another horizontal line; see lower part of Fig. 9. Marking of the scale is then effected by small bright dots which must be in alignment with the lines on the scale. The dots indicating the distortion values have a greater brightness. It is now possible correctly to associate the distortion value with the respective signal element. This equipment ensures a reading error of not more than 1/2%.

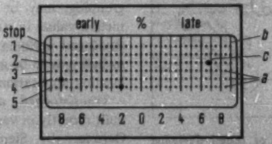
The equipment described is quite large, and only a limited number were built for use in laboratories. However, a smaller teleprinter distortion meter, operating on the same principle, will soon be introduced on the market, but the crystal in the new meter is replaced by an *L-C* (inductance-capacitance) oscillator which can be adapted to the speed of the teleprinter.

The speed deviation meter, Fig. 10, fills another need. It evaluates the incoming signals and determines the average deviation from the nominal speed. Signals must be sent by an autotransmitter at the full telegraph speed. The deviation meter contains a bridge circuit with three arms formed by resistors, while the fourth arm contains a capacitor, charged and discharged by the armature of a relay; see Fig. 11. The current in this arm is consequently determined by the frequency of the armature operations. A switch with relay and a resonant circuit of about 4 cps insures that the relay of the measuring bridge is actuated only by the start pulse transitions of the signals, transmitted in uninterrupted succession. Thus, the mean value of the current in this bridge arm depends directly on the telegraph speed. Irregularities caused by distortion of the individual pulse

Screen of precision start-stop distortion meter



All transitions together, range ±50%



Transitions split up vertically, range ±10%

Fig. 9. Display of high-precision distortion meter

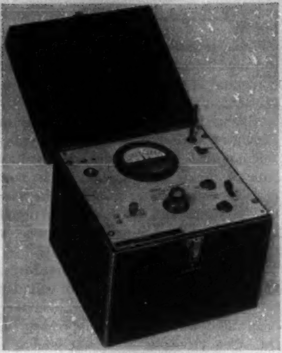


Fig. 10. Speed deviation meter

transitions are of no consequence because this is an integrating instrument.

One of the most rewarding applications is the verification of subscriber station apparatus in a Telex (International Teleprinter Exchange) network. The deviation meter is connected to the line terminating circuit in the telegraph office. The machine in the subscriber station need not be attended. It is switched on by remote control and its station identification code is requested by tripping its answerback unit. While the 20 signals of the answerback code are received, the deviation meter adjusts itself to the required degree of accuracy. Distortion will not cause an undue measuring error. Assuming a distortion of 10%, the measuring error cannot exceed 0.07%. According to CCITT recommendations, a speed deviation of 0.75% is permissible.[15]

All this equipment is for checking and adjusting teleprinter apparatus, line sections, and complete operational teleprinter circuits. Correct distortion measurements are possible only on the polar-current-operated side of the circuits.

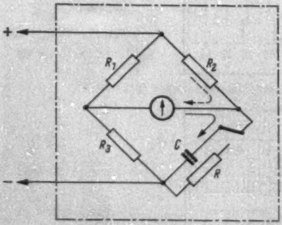


Fig. 11. Bridge circuit of speed deviation meter

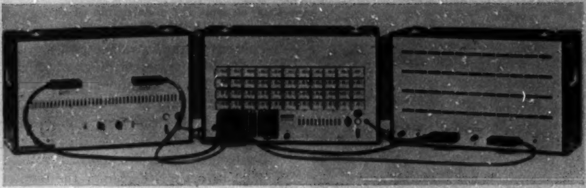


Fig. 12. Distortion analyzer (Siemens and Halske)

Hence, in the terminal and intermediate transmission offices, all signals are converted into their polar form so that measurements at these points will not involve any difficulties. In the neutral current operated subscriber station loops, however, which may extend over a number of miles, the signal shape is affected by the capacitance of the line and the inductance introduced by receive magnets, relays, etc.

To permit distortion measurements, the operate point of each relay and receiver would have to be known. In view of this difficulty, the CCITT decided to base all measurements, as well as recommended values, on polar current. The subscriber station line, plus neutral–polar conversion relay, are considered as forming part of the apparatus. The rule, stating that the total distortion of a subscriber station apparatus must not exceed 10%, therefore implies that the indicated value covers the apparatus, together with subscriber station line and converter relay.[16] This portion of an over-all circuit is referred to as local end.

Distortion Analyzers

A rather big undertaking remains; that of establishing an interrelationship between signal distortion, receive margin of terminal apparatus, and error rate in reception, as well as providing supporting data needed to plan extensive teleprinter networks.

Distortion analyzers are required for recording not only the peak values of distortion but each and every distortion of all pulse transitions. The first analyzer, probably built by the British Post Office,[17] had 20 counters which, in an optional grouping, may be assigned to the distortion values in steps of 1%. As it was developed in the vacuum-tube era, it is so bulky as to be suitable only for fixed-location applications in laboratories.

Transistorized analyzers have been built recently in various countries, one in Germany by Siemens and Halske.[18,19] This analyzer is accommodated in three carrying cases; see Fig. 12. It includes 40 counters which may be used either in steps of 2.5% over the entire range of ±50% or in increments of 1% over a range of ±20% distortion. In the latter case, a further counter will reveal whether and how often a distortion exceeding 20% was present in the line.

Furthermore, the measured signals and pulse transitions are counted. The count may be stopped automatically after 10^3 or 10^4 transitions to facilitate evaluation.

The French PTT (Post, Telegraph, Telephone Administration) has built an analyzer which permits measurements increments of 0.5% over a range of 25%. Some examples of measurements and the results obtained with such analyzers are given herewith.

The British Post Office has made a considerable number of measurements on circuits which consisted of $n = 1 \ldots 6$ sections and has drawn up approximative equations for the mean values and the stray of the distortion values in dependence of n.[20]

Fig. 13 shows a distribution curve based on measurements published by the French PTT.[21] The close agreement with a Gaussian distribution is immediately apparent. Also, these measurements justify the conclusion that the distortion of a multisection line can be calculated in advance with the aid of a square law,

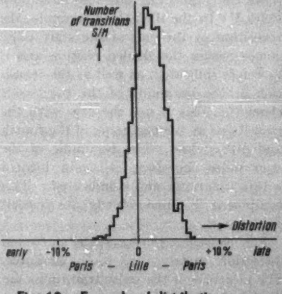


Fig. 13. Example of distribution curve

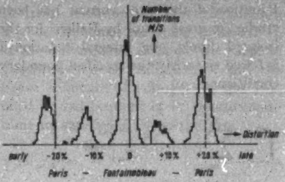


Fig. 14 (left). Divided distribution curve

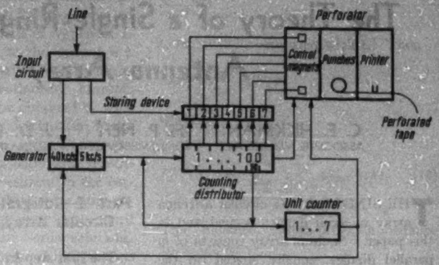


Fig. 15 (right). Analyzer with perforated tape

corresponding roughly to the equation given in the previous paragraph.

Registration of not more than 500 transitions will afford a satisfactory picture of how distortion values are distributed. A practical suggestion at this point is to mark the width of the distribution curve by those values which are exceeded with a probability of 1 to 1,000. This is a material change from the figure 1 to 10^6 as used for defining the degree of conventional distortion, but the substituted value corresponds more closely to practical requirements.

The distribution curves have clearly shown the influence of characteristic distortion, which is a function of the length of the shortest signal elements. Fig. 14 depicts an extreme case.[22] Signals distorted by 20% alternately in a leading and a lagging sense were sent, together with interposed undistorted signals, over two v-f transmission sections by a distorted-signal generator. Only the mark-to-space transitions were recorded. The diagram is divided into five individual distribution curves while only three were to be expected: leading, undistorted, and lagging. Undistorted transitions behind transitions with lagging distortion resulted in signals having elements of reduced length. The characteristic distortion of the line prolonged these signal elements so that part of the signal elements transmitted undistorted appears as a separate distribution curve in the range of 10% lagging. Analogously, the curve at 10% leading stems from the transitions which had a 20% leading distortion.

In practical operation, no individual and discrete distortion values are present at the beginning of a line but, in accordance with the distribution curve, occur only between two limits. The final result is, therefore, not as extreme as shown by the curves in Fig. 14, although an appreciable deformation will result. An analyzer, therefore, should cover not only the distribution curves of the pulse transitions but also the position of all individual transitions so as to obtain information on the length of individual pulses and the changes in shape to which they are subjected from one line section to the other.

A suitable arrangement has been proposed to the CCITT by the author.[23] The operating principle is based on an electronic counting chain, Fig. 15, in which each 1%-section of a signal element is assigned to one of the 7-unit code combinations, this code affording 128 permutations. For each pulse transition, a high-speed punch—rated for a minimum of 50 characters per second—will punch the respective code combination into paper tape. The code perforations can be supplemented so that the teleprinter signal-proper becomes recognizable.

Equipment of this type is not bulky, can be easily transported and, compared with previous analyzers, is not expensive. Several such analyzers could be connected at different points of a communication path for simultaneous registration of distortion values. Evaluation of the punched information could be accomplished at a central location. The simplest method would require a single counter which, with repeated feed-through of the tape, would count the frequency of the individual 7-unit code combinations; that is, the distortion values, thereby furnishing the curve for distribution. Even more complicated interpreting procedures need not be alarming because, with the electronic computer, they should not prove unduly difficult.

Conclusions

The CCITT-recommended standards are of practical value. Nevertheless, great improvements need to be made in the future.

References

1. List of Definitions of Essential Telecommunication Terms, Part I. International Telecommunications Union, Geneva, Switzerland, June 1957.

2. Recommendations in: Documents of the VIIIth Plenary Assembly of the CCITT, 1956. International Telecommunication Union, Feb. 1957.

3. *Documents* of Committees 8 and 9, 1957–60. International Telegraph and Telephone Consultative Committee.

4. Measurement of Telegraph Transmission, H. Nyquist, R. B. Shanck, S. I. Cory. *AIEE Transactions*, vol. 46, Feb. 1927, pp. 367–76.

5. Unité de measure pour vérifier la qualité de circuits télégraphiques, A. Jipp. *Documents*, Deuxième Réunion, International Telegraph Consultative Committee, Berne, Switzerland, Bureau de l'Union Internationale der Télécommunication, Oct. 1929.

6. Verzerrungsmesser für Telegraphie, A. Jipp, U. O. Römer. *Telegraphen-Fernsprech- und Fernsehtechnik*, Berlin, Germany, May 1932.

7. *Document 55.* Committee 9, definition 33.08, p. 8. (Part of reference 3.)

8. Recommendation *B 51.*

9. See reference 7, p. 20.

10. See reference 7, p. 125.

11. *Document 36.* Committee 9, p. 3. (Part of reference 3.)

12. See reference 7, p. 100.

13. Die Messgeräte der Telegraphie, G. Keller. *Elektrotechnische Zeitschrift*, Wuppertal-Elberfeld, Germany, 1939.

14. Der Fernschreib-Verzerrungsmelder, ein automatisches Prüfgerät, E. Schenk. *Nachrichtentechnische Zeitschrift*, Brunswick, Germany, vol. 12, 1959, pp. 609–12.

15. Recommendation *C 4.*

16. *Document 62.* Committee 9, p. 39. (Part of reference 3.)

17. A Telegraph Distortion Analyzer, L. K. Wheeler, A. C. Frost. *The Post Office Electrical Engineers Journal*, London, England, Apr. 1954.

18. Neue Messgeräte der Fernschreibtechnik, G. Keller, D. von Sanden. *Nachrichtentechnische Zeitschrift*, 1958, pp. 411 ff.

19. Ein Fernschreib-Verzerrungsspektrometer, G. Funk. *Ibid.*, 1960, p. 115.

20. *Document 24.* Committee 9, pp. 9 ff. (Part of reference 3.)

21. *Document 46.* Committee 9, p. 31. (Part of reference 3.)

22. *Ibid.*, p. 36.

23. *Document 7.* Committee 8. (Part of reference 3.)

The Theory of a Single-Ring Circular Antenna Array

C. E. HICKMAN
ASSOCIATE MEMBER AIEE

H. P. NEFF
NONMEMBER AIEE

J. D. TILLMAN
ASSOCIATE MEMBER AIEE

THE SINGLE-RING circular antenna array of axial dipoles is analyzed in this paper. Such an array consists of m parallel dipoles, equally spaced on the circumference of a circle whose radius is ρ. The two principal properties with which a designer is concerned are (1) the radiation pattern and (2) the impedance of each element in an array, both of which can be found if the current distribution along the element is known. To synthesize a specified radiation pattern, the voltages at the center of each element must be determined.

The first section of the paper shows what currents are possible on a circular array; the second, what patterns result from these currents; and the third, how to obtain a specified pattern. The circular array was first studied by Stenzel,[1] and later by Knudsen;[2] the first synthesis technique by DuHamel.[3] None of these analyses touched upon impedances. The current distribution and impedance problem was examined by King,[4] and on his work our Part I is based. The general methods of DuHamel are then applied in Parts II and III to the analysis of Part I.

The circular antenna array has one significant advantage over the widely used linear array in that, if a satisfactory directional pattern can be obtained, the beam can be pointed in different directions by simply shifting the excitation around the ring. The demand for steerable beams has increased greatly in recent years but a large linear antenna cannot be moved physically to point a beam in the required direction. With linear arrays, its shape is degraded if the beam is pointed at angles larger than about 45 degrees off normal to the array. A circular array provides for such applications.

Part I—Integral Equations for a Circular Array

The first step in analyzing circular antenna arrays is determination of possible current sets on elements of the array. Any desired set of terminal voltages can be used, and the resulting currents must be such that the electric fields satisfy the boundary conditions on all surfaces of the elements.

Current distribution is conveniently ascertained by using a cylindrical coordinate system with the z-axis lying along the axis of one dipole, and $z=0$ at its center. Centers of all the other dipoles also lie in the plane $z=0$. This is shown in Fig. 1, which also gives the meaning of several symbols used in the analysis. Initial assumptions are:

1. $\beta a << 1$. If this is the case, there will be only a z-component of current in each antenna.
2. The end faces of the antennas are neglected, the current there being taken as zero.

If the elements are assumed to be perfect conductors, the boundary condition which must be satisfied is that the electric field tangential to each element on its cylindrical surface must be zero. While this assumption is not necessary, it greatly simplifies the analysis, and for a good conductor it leads to no significant errors.

Each dipole of the array is driven by a voltage V_i where the index i runs from 1 to m and identifies each element of the array. How to determine the axial distribution of the total current has been studied in great detail by Hallen[5] for the isolated dipole, and integral equation 1 is found to satisfy the specified boundary conditions.

$$\int_{-h}^{h} \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} I_{zi}(\bar{z}) \frac{e^{-j\beta R_{mi}}}{R_{mi}} d\bar{z} = \frac{-j4\pi}{\eta}\left(C_m \cos \beta z + \frac{V_m}{2} \sin \beta |z|\right) \qquad (1)$$

where

$$\eta = \sqrt{\frac{\mu}{\epsilon}}$$

is the intrinsic impedance of free space, and

$$R_{mi} = \sqrt{(z-\bar{z})^2 + d_{mi}^2}$$

where d_{mi} is the distance from the ith element-to-element number m. If $i=m$

$$R_{mm} = \sqrt{(z-\bar{z})^2 + a^2}$$

Equation 1 is developed in Appendix I, and in the summation, the limits on the index i are for m odd. If m is even, i runs from $-(m/2)+1$ to $m/2$. In this paper, the limits will be written for m odd.

These values of R_{mi} assume that the vector potential is the same as if the currents were filaments of zero thickness along the axis of each element. The currents on the circular array of m elements are thus found to be related to the m base voltages by m simultaneous integral equations similar to equation 1.

The solution of such equations has not been accomplished, even for the case of $m=2$. In analyzing a linear array, a number of simplifying assumptions must be made at this point, but these are neither necessary nor desirable for the circular array. The m simultaneous integral equations are identical if every antenna is in the same total environment, and the geometrical situation of each

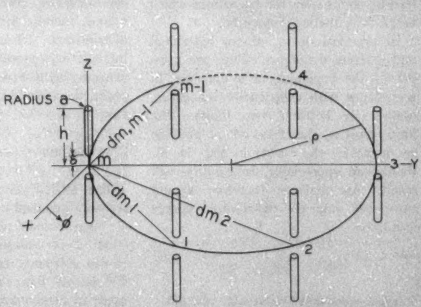


Fig. 1. Co-ordinate system used for determining currents and impedances

---

Paper 61-61, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 2, 1960; made available for printing December 22, 1960.

C. E. Hickman, H. P. Neff, and J. D. Tillman are with the University of Tennessee, Knoxville, Tenn.

The work reported in this paper was supported by the Air Force Cambridge Research Center under Contract *AF 19(604)-4967*.

antenna is already the same. If the electrical environment of each antenna is to be the same, the scalar potential V_i at the ith antenna must be the same in magnitude for all values of i, and the phase difference of voltages V_i and V_{i+1} must be independent of i. These conditions are satisfied if V_i has the form

$$V_i^{(n)}=V^{(n)}e^{\frac{j2\pi ni}{m}} \tag{2}$$

where n is any integer and $V^{(n)}$ is a constant.

If V_i has the form given by equation 2, the current distribution along each element must be the same, although differing in phase from the current in an adjacent element by $(2\pi/m)n$. If V_i is given by equation 2, then

$$I_{zi}(\bar{z})=I_z^{(n)}(\bar{z})e^{\frac{j2\pi ni}{m}}=I_z^{(n)}f^{(n)}(\bar{z})e^{\frac{j2\pi ni}{m}} \tag{3}$$

The function $I_z^{(n)}(z)$ is a complex function of z which is independent of i. Since the summation of equation 1 is independent of any general element k, the integral equation becomes

$$\int_{-h}^{h} I_z^{(n)}(\bar{z})K^{(n)}(z,\bar{z})d\bar{z} =\frac{-j4\pi}{\eta}\left(C_n^{(n)}\cos\beta z+\frac{V^{(n)}}{2}\sin\beta|z|\right) \tag{4}$$

where

$$K^{(n)}(z,\bar{z})=\sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{\frac{j2\pi ni}{m}}\frac{e^{-j\beta R_{mi}}}{R_{mi}}$$

If the function $V^{(n)}$ is chosen according to the equation

$$V^{(n)}=\frac{1}{m}\sum_{i=1}^{m} V_i e^{\frac{-j2\pi ni}{m}} \tag{5}$$

then any set of m voltages $V^{(i)}$ can be written as the sum of m sets of voltages $V^{(n)}$. The sets $V_i^{(n)}$ are called the sequence components of V_i, and $V_i^{(n)}$ is called the nth sequence voltage. This resolution into sequence components is, of course, identical to the resolution into symmetrical components used in polyphase circuit analysis.

For an arbitrary set of base voltages V_i applied to a circular array, equation 5 is used to find the m sequence sets $V_i^{(n)}$. Each of these gives an integral equation of the form of equation 4, so that instead of m simultaneous integral equations, the result is m uncoupled integral equations, each of which can be solved approximately. The superposition of these solutions gives the actual currents on the elements of the array.

The solution of equation 4 can be obtained by a process of iteration, due to King,[4] and it is sketched in Appendix I. A first approximation to the current is a solution, composed of trigonometric terms only, called the zeroth-order current. The true current can be obtained by adding a correction term which is a function of the true current. An approximate correction results from substituting zeroth-order current for true current in the correction term, giving a first-order approximation. This process may be carried out again to obtain second-order solutions, or solutions of any order, but integrations for higher-order solutions become almost impossible to carry out.

$$[I_z^{(n)}(z)]_1=\frac{j2\pi V^{(n)}}{\eta\psi^{(n)}}\,\frac{M_0(z)+\dfrac{M_1^{(n)}(z)}{\psi^{(n)}}}{F_0(h)+\dfrac{F_1^{(n)}(h)}{\psi^{(n)}}} \tag{6}$$

The functions used in equation 6 are defined in Appendix II. The first-order impedance is found by setting $z=0$, which gives the terminal current, and then solving for the ratio $V^{(n)}/I_z^{(n)}(0)$, which is the terminal impedance.

The first-order impedance is then

$$Z_1^{(n)}=\frac{-j\eta\psi^{(n)}}{2\pi}\,\frac{\cos\beta h+\dfrac{F_1^{(n)}(h)}{\psi^{(n)}}}{\sin\beta h+\dfrac{M_1^{(n)}(0)}{\psi^{(n)}}} \tag{7}$$

The simpler and less accurate zeroth-order impedance is

$$Z_0^{(n)}=\frac{j\eta\psi^{(n)}}{2\pi}\cot\beta h \tag{8}$$

This is a poor approximation, being a pure reactance. Physically, it is closely related to making the assumption that the antenna is an m-wire open-circuit transmission line. If $h=\lambda/4$, the first-order impedance becomes

$$Z_1^{(n)}=\frac{j\eta\psi^{(n)}}{2\pi}\,\frac{\dfrac{F_1^{(n)}(h)}{\psi^{(n)}}}{1+\dfrac{M_1^{(n)}(0)}{\psi^{(n)}}} \tag{9}$$

An approximation called the modified zeroth-order solution is obtained by considering the ratio

$$\frac{M_1^{(n)}(0)}{\psi^{(n)}}$$

to be small as compared with 1. Then

$$Z_{01}^{(n)}=\frac{-j\eta}{2\pi}F_1^{(n)}(h)$$

where

$$F_1^{(n)}(h)=-\sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{\frac{j2\pi ni}{m}}C_{d_{mi}}(h,h) \tag{10}$$

The modified zeroth-order impedance then becomes

$$Z_{01}^{(n)}=\frac{j\eta}{2\pi}\sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{\frac{j2\pi ni}{m}}C_{d_{mi}}(h,h) \tag{11}$$

The function $C_d(h,h)$ given in Appendix II may be expressed in terms of the generalized sine and cosine integral functions, for which extensive tables are available. It recently has been tabulated by Mack.[6] An extensive set of tables of the modified zeroth-order impedance has been compiled by Hickman,[8] for m from 2 to 20, and $\beta\rho$ from 0 to 12. Tables I and II give a few values therefrom.

The impedance of any element can now be found from the sequence impedance. The voltage at the center of the ith element is

$$V_i=\sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} V^{(n)}e^{\frac{j2\pi ni}{m}} =\sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} I^{(n)}Z^{(n)}e^{\frac{j2\pi ni}{m}} \tag{12}$$

The current is

$$I_i=\sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} I^{(n)}e^{\frac{j2\pi ni}{m}} \tag{13}$$

so the impedance is

Table I. Modified Zeroth-Order Sequence Resistance for Circular Antenna Arrays of Half-Wave Dipoles

m	n	$\beta\rho=0$	$\beta\rho=1$	$\beta\rho=2$	$\beta\rho=3$	$\beta\rho=4$
4	0	292.5	190.5	40.2	21.5	72.0
4	1	0.0	48.2	98.1	74.1	50.3
4	2	0.0	5.6	56.2	122.8	101.8
6	0	438.8	285.9	59.7	22.7	62.0
6	1	0.0	72.2	142.2	81.0	21.9
6	2	0.0	4.2	42.5	96.9	99.4
6	3	0.0	0.1	9.8	60.3	134.2

Table II. Modified Zeroth-Order Sequence Reactance for Circular Antenna Arrays of Half-Wave Dipoles

m	n	$\beta\rho=0$	$\beta\rho=1$	$\beta\rho=2$	$\beta\rho=3$	$\beta\rho=4$
4	0	170.2	41.6	−37.7	53.5	75.0
4	1	0.0	78.4	52.6	23.6	47.2
4	2	0.0	55.0	102.0	69.5	0.8
6	0	255.3	−76.1	−85.7	26.0	39.8
6	1	0.0	100.8	40.0	−23.4	26.1
6	2	0.0	48.0	91.6	79.2	37.0
6	3	0.0	33.7	77.9	117.6	89.3

$$Z_1=\frac{\displaystyle\sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} I^{(n)}Z^{(n)}e^{\frac{j2\pi ni}{m}}}{\displaystyle\sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} I^{(n)}e^{\frac{j2\pi ni}{m}}}$$

Part II—Patterns of Sequence Currents

Part I demonstrated that currents on a circular array can be resolved into sequence sets. Next in order is to determine the radiation pattern of such a set. The pattern for arbitrary excitation can then be found by resolving the base voltages into sequence components. The sequence currents are found by use of the sequence impedances; the sequence patterns are found from the currents; and finally these are superimposed to give the actual pattern.

A spherical co-ordinate system shown in Fig. 2 is desirable for helping to find the radiation pattern. The origin is at the center of the circle and the direction $\phi=0$ lies through the mth element. A point in space has co-ordinates r, θ, ϕ (spherical); a point on one of the dipoles has co-ordinates ρ, ϕ, z (cylindrical). Equation 3 showed that for the nth sequence, the current on the ith element is

$$I_i=I^{(n)}f^{(n)}(\bar{z})e^{\frac{j2\pi ni}{m}} \qquad (14)$$

If h/a is large, $f^{(n)}(\bar{z})$ is well-approximated by the zeroth-order term $\sin \beta(h-|\bar{z}|)$.

The vector potential **A** at some point $(\mathbf{r}, \theta, \phi)$ is given by

$$\mathbf{A}(r, \theta, \phi)=\frac{\mu}{4\pi}\int \mathbf{J}(\rho, \bar{\Phi}, \bar{z})\frac{e^{-j\beta R}}{R}\,d\bar{v} \qquad (15)$$

in which R is the distance from $(\rho, \bar{\Phi}, \bar{z})$ to (r, θ, ϕ), and integration is carried out through the entire region in which there are currents. This gives

$$A_z(r, \theta, \phi)=\frac{\mu I^{(n)}}{4\pi}\sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{\frac{j2\pi ni}{m}}\times \int_{\bar{z}=-h}^{\bar{z}=h} f^{(n)}(\bar{z})\frac{e^{j\beta R_i}}{R_i}\,d\bar{z} \qquad (16)$$

R_i is the distance from the point $(\rho, 2\pi i/m, \bar{z})$ on the ith element to the point (r, θ, ϕ). There is only a z-component of A, since there are only z-components of J. Provided $R_i>>4\rho^2/\lambda$, so that rays from $(\rho, 2\pi i/m)$ to (r, θ, ϕ) are essentially parallel,

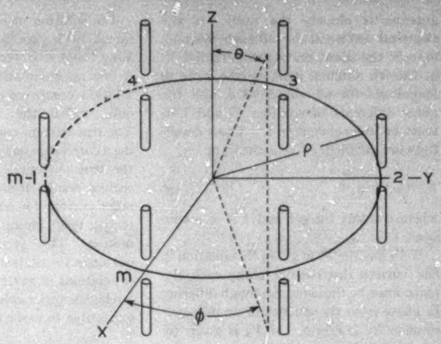


Fig. 2. Co-ordinate system used in calculation of radiation pattern

$$R_i=r-\beta\rho\sin\theta\cos\left(\phi-\frac{2\pi}{m}i\right)-\beta\bar{z}\cos\theta \qquad (17)$$

The factor R_i in the denominator of equation 16 can be approximated by r. Then

$$A_z(r, \theta, \phi)=\frac{\mu I^{(n)}}{4\pi}\frac{e^{-j\beta r}}{r}\times \int_{-h}^{h} f^{(n)}(\bar{z})e^{j\beta\bar{z}\cos\theta}\,d\bar{z}\times \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{j\left[\frac{2\pi ni}{m}+\beta\rho\sin\theta\cos\left(\phi-\frac{2\pi i}{m}\right)\right]} \qquad (18)$$

The electric field has only a θ component

$$E_\theta(r, \theta, \phi)=-j\omega A_\theta=-j\omega A_z\sin\theta \qquad (19)$$

The integral in equation 18 leads to what is usually called the element pattern factor. It is

$$F_e^{(n)}=\frac{\beta}{2}\int_{-h}^{h} f^{(n)}(\bar{z})\sin\theta\, e^{j\beta\bar{z}\sin\theta}\,d\bar{z} \qquad (20)$$

Designating the sum in equation 18 by $S^{(n)}(\theta, \phi)$, the electric field is

$$E_\theta(r, \theta, \phi)=\frac{j\omega\mu}{4\pi}I^{(n)}F_e^{(n)}(\theta)S^{(n)}(\theta, \phi)\frac{e^{-j\beta r}}{r} \qquad (21)$$

The problem is thus reduced to the evaluation of $F_e^{(n)}(\theta)$ and $S^{(n)}(\theta, \phi)$.

If $f^{(n)}(\bar{z})$ can be approximated by $\sin \beta(h-|z|)$, the integral in equation 20 is well-known, and for any n

$$F_e(\theta)=\frac{\cos(\beta h\cos\theta)-\cos\beta h}{\sin\theta} \qquad (22)$$

The validity of this approximation for use in pattern computation will be considered later. The sum $S^{(n)}(\theta, \phi)$ has been investigated by Stenzel[1] for $n=0$, by Chireix[7] for n even, and by Knudsen[8] for n even or odd—although his analysis is not available in English. For any n, Tillman et al.[9] found

$$S^{(n)}(\theta, \phi)=me^{jn\phi}J_n(\beta\rho\sin\theta)+ m\sum_{p=1}^{\infty}\left[j^{(mp-n)}e^{-j(mp-n)\phi}J_{mp-n}(\beta\rho\sin\theta)+j^{(mp+n)}e^{j(mp+n)\phi}J_{mp+n}(\beta\rho\sin\theta)\right] \qquad (23)$$

The electric field is then found by substituting equations 22 and 23 into 21:

$$E_\theta(r, \theta, \phi)=j\frac{\eta}{2\pi}\times I^{(n)}m\frac{\cos(\beta h\cos\theta)-\cos\beta h}{\sin\theta}\frac{e^{-j\beta r}}{r}\times \left\{j^n e^{jn\phi}J_n(\beta\rho\sin\theta)+ \sum_{p=1}^{\infty}\left[j^{mp-n}e^{-j(mp-n)\phi}J_{mp-n}(\beta\rho\sin\theta)+j^{(mp+n)}e^{j(mp+n)\phi}\times J_{mp+n}(\beta\rho\sin\theta)\right]\right\} \qquad (24)$$

Equation 24 is somewhat formidable, but in most practical cases some simplifying approximations can be made, since if $k>x$, $J_k(x)$ is small. The largest value of the argument of the Bessel functions in 24 is for $\theta=\pi/2$ and is $\beta\rho$. The usual design procedure is to choose the circumference $\beta\rho$ and the number of elements m so that $J_n(\beta\rho)$ is not negligible, but $J_{m-n}(\beta\rho)$ is small. This requires $\beta\rho\geq n$, and $m-n>>\beta\rho$. This condition is easily satisfied if $\beta\rho\approx n$, and $m>\beta\rho+n+3$. If these conditions are satisfied, all terms in the summation are small, and are called residuals. The leading term is called the principal term. The electric field is thus

$$E_\theta(r, \theta, \phi)=j\frac{\eta}{2\pi}mI^{(n)}F_e^{(n)}(\theta)j^nJ_n(\beta\rho\sin\theta)e^{jn\phi}+\text{residuals} \qquad (25)$$

The justification for approximating $F_e^{(n)}(\theta)$ by $[\cos(\beta h\sin\theta)-\cos\beta h]/\sin\theta$

can now be considered. King[4] has shown, for the isolated element, the departure of the true function from this approximation occurs mainly for θ near zero. But $J_n(\beta\rho \sin\theta)$ is small for θ near zero if $n \neq 0$, so the errors occur only in the region where the field is small. It follows that equation 22 is an acceptable approximation if $n \neq 0$. The error for the case of $n = 0$ is larger, but usually acceptable.

Part III—Synthesis of a Desired Pattern

The remaining problem is how to combine the sequence patterns of the form given by equation 25 in a manner which will give a desired pattern. The sequence number n can take on values

$$-\frac{m-1}{2} \leq n \leq \frac{m-1}{2},\ m \text{ odd}$$

or

$$\left(-\left(\frac{m}{2}-1\right) \leq n \leq \frac{m}{2},\ m \text{ even}\right)$$

If the residuals are to be held small enough so that they may be neglected, the circumference must be determined from the largest value of n. If this is done, the variation of the pattern with ϕ will be of the general form

$$\sum_{k=-\infty}^{\infty} A_k e^{jk\phi}$$

and only those values of A_k for which $-(m-1)/2 < k < (m-1)/2$ will be of appreciable size. These values of A_k can be related to the sequence currents, as subsequently explained. An arbitrary set of currents can be expressed as a sum of the sequence currents by

$$I_i = \sum_{n=-\frac{m-1}{2}}^{\frac{m-1}{2}} I^{(n)} e^{\frac{j2\pi ni}{m}} \tag{26}$$

From equation 25, neglecting residuals, the electric field due to each term of 26 is

$$E_\theta{}^{(n)}(r,\theta,\phi) = j\frac{\eta}{2\pi} mI^{(n)} \frac{e^{-j\beta r}}{r} F_e{}^{(n)}(\theta)\, j^n e^{jn\phi} J_n(\beta\rho \sin\theta) \tag{27}$$

Table III. Currents and Impedances for 6-Element Circular Array

Element No.	Current	Impedance
6	$1.49\,\underline{/-62.0}$	$173.0-j49.0$
1,5	$2.75\,\underline{/-13.7}$	$67.3-j25.0$
2,4	$2.75\,\underline{/\ 13.7}$	$25.2-j10.4$
3	$1.49\,\underline{/\ 62.0}$	$-28.8-j25.0$

It is customary to use an array pattern factor, $F^{(n)}(\theta,\phi)$, letting

$$F^{(n)}(\theta,\phi) = \frac{E_\theta{}^{(n)}(r,\theta,\phi)r}{j\frac{\eta}{2\pi} me^{-j\beta r}} = j^n I^{(n)} J_n(\beta\rho \sin\theta) e^{jn\phi} F_e{}^{(n)}(\theta) \tag{28}$$

The pattern factor for an arbitrary set of currents is found by summing with respect to n over all the possible sequences, giving

$$F(\theta,\phi) = \sum_{n=-\frac{m-1}{2}}^{n=\frac{m-1}{2}} j^n I^{(n)} J_n(\beta\rho \sin\theta) e^{jn\phi} F_e{}^{(n)}(\theta) \tag{29}$$

If the zeroth-order approximation

$$F_e{}^{(n)}(\theta) = F_e(\theta) = \frac{\cos(\beta h \cos\theta) - \cos\beta h}{\sin\theta}$$

is sufficiently accurate, equation 29 becomes

$$F(\theta,\phi) = F_e(\theta) \sum_{n=-\frac{m-1}{2}}^{n=\frac{m-1}{2}} j^n I^{(n)} J_n(\beta\rho \sin\theta) e^{jn\phi} \tag{30}$$

Equation 30 gives the pattern of any set of currents in a single-ring array. The quantity in the summation is called the array factor $F_a(\theta,\phi)$.

The general synthesis technique is now evident. At a specified angle θ_0, let the required pattern be $f(\phi)$. The first step is to expand $f(\phi)$ in an exponential Fourier series

$$f(\phi) = \sum_{k=-\infty}^{\infty} A_k e^{jk\phi} \tag{31}$$

in which A_k is, as usual,

$$A_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\phi) e^{-jk\phi} d\phi$$

It is next necessary to select a number of terms K, so that $f(\phi)$ is well approximated by

$$f(\phi) \approx \sum_{k=-K}^{K} A_k e^{jk\phi} \tag{32}$$

K is then equated to $(m-1)/2$, so that the number of elements is

$$m = 2K+1 \tag{33}$$

The sequence currents $I^{(k)}$ are now found by equating $j^k I^{(k)} J_k(\beta\rho \sin\phi_0)$ to $A^{(k)}$ so that

$$I^{(k)} = \frac{A^{(k)}}{j^k J_k(\beta\rho \sin\theta_0)} \tag{34}$$

The element currents are then

$$I_i = \sum_{k=-K}^{K} I^{(k)} e^{\frac{2j\pi ki}{m}} \tag{35}$$

From equations 28 and 30, the far-zone electric field at $\theta = \theta_0$, is thus

$$E_\theta(r,\theta,\phi) = j\frac{\eta}{2\pi}(2K-1)\frac{e^{-j\beta r}}{r} f(\phi) F_e(\theta_0) \tag{36}$$

and the azimuthal variation is $f(\phi)$, as desired.

Since the kth sequence voltage is $V^{(k)} = I^{(k)}Z^{(k)}$, the terminal voltage at the ith element is

$$V_i = \sum_{k=-K}^{K} I^{(k)} Z^{(k)} e^{\frac{j2\pi ki}{m}} \tag{37}$$

and the impedance of the ith element is

$$Z_i = \frac{V_i}{I_i} = \frac{\sum_{k=-K}^{K} I^{(k)} Z^{(k)} e^{\frac{j2\pi ki}{m}}}{\sum_{k=-K}^{K} I^{(k)} e^{\frac{j2\pi ki}{m}}} \tag{38}$$

The element impedance is thus related to the pattern by way of equations 34, 38, and 11.

There are two sources of error in this synthesis technique. One is the approximation of $f(\phi)$ by a truncated Fourier series in equation 32, and the other is the neglect of residuals. For most purposes, the errors can be made so small as to be of no consequence.

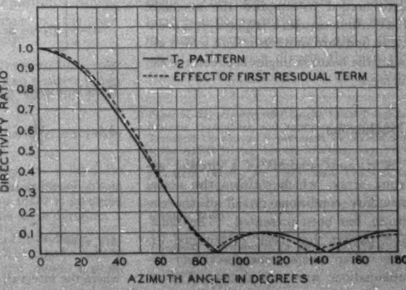


Fig. 3. Pattern of 6-element single-ring array

Example

Suppose the problem is to synthesize a circular array which will have a directional pattern with side lobes 20 decibels below the main beam, and a beam width of about 70 degrees. Employing the method of DuHamel[3] using Tchebyscheff polynomials, a satisfactory expression for $f(\phi)$ is

$$f(\phi)=2.70+4.5\cos\phi+2.8\cos 2\phi \qquad (39)$$

The coefficients A_k are then: $A_0=2.70$, $A_1=A_{-1}=2.25$, and $A_2=A_{-2}=1.4$. Since $K=2$, $m=5$. The circumference $\beta\rho=K=2$, and $\rho=2\lambda/2\pi=0.318\lambda$ is satisfactory. The first residual is then

$$J_{m-2}(2)=J_3(2)=0.129$$

and since the coefficient for A_2 results from $J_2(2)=0.353$, the first residual is not small enough to neglect. If the number of elements is increased to $m=6$, the first residual is proportional to $J_4(2)=0.034$, which is small. Equations 34 and 35 are now used to compute the currents and 38 and 11 to find the impedance. The element currents and impedances found in this way are given in Table III. The pattern, including the effect of the first residual, is shown in Fig. 3.

This example is rather simple, but it serves to illustrate the method, and certain results are of interest. The pattern is of an acceptable shape, with a beamwidth of about 80 degrees, which is the narrowest obtainable with this size aperture without entering the supergain region. Since the array has six elements, the beam can be directed to any of six pointings by simple switching methods, so that this method of steering the beam results in 60-degree increments in pointing. The beam width of 80 degrees thus agrees well with steering increment, so that small overlap of half-power points results when the beam is steered. This is a rather general property of circular arrays. Smaller increments in steering can be obtained by using more elements than needed to suppress the residuals. If a continuous scan is required, it is possible to find current distributions which will point the beam to angles between the elements.[10]

Conclusions

A rigorous analysis of a circular antenna array, which satisfies the actual boundary conditions on all of the elements, has been presented. The radiation pattern has been determined exactly, after which some customary approximations were introduced. The results of the pattern analysis lead to a simple synthesis method, which will approximate a desired pattern with good accuracy.

Appendix I

Development of integral equation 1 of the paper is outlined here. The boundary condition which must be met on the surface of element 1 is that the tangential electric field be zero. That is

$$E_z(a,z)=0 \qquad (-h\le z\le h) \qquad (40)$$

At any point in space, the electric field is

$$\mathbf{E}=-\nabla\phi-j\omega\mathbf{A} \qquad (41)$$

where ϕ and $\mathbf{A}$ are the scalar and vector potentials respectively. They are related by

$$\nabla\cdot\mathbf{A}=-j\frac{\omega}{c^2}\phi \qquad (42)$$

Then substituting 42 into 41

$$\mathbf{E}=\frac{c^2}{j\omega}\left[\nabla(\nabla\cdot\mathbf{A})+\frac{\omega^2}{c^2}\mathbf{A}\right] \qquad (43)$$

Since there are only z-components of current on the antennas, $\mathbf{A}$ will have only a z-component. $E_z(r,z)$ is then

$$E_z(r,z)=-\frac{j\omega}{\beta^2}\left(\Delta\frac{\partial^2 A_z}{\partial z^2}+\beta^2 A_z\right) \qquad (44)$$

Since this must be zero when $r=a$

$$\left(\frac{\partial^2 A_z}{\partial z^2}+\beta^2 A_z=0\right)_{r=a} \qquad (-h\le z\le h) \qquad (45)$$

The solution of this is

$$A_z(a,z)=-\frac{j}{c}(C_1\cos\beta z+C_2\sin\beta z) \qquad (\delta\le z\le h)$$

$$=-\frac{j}{c}(C_3\cos\beta z+C_4\sin\beta z) \qquad (-h\le z\le -\delta) \qquad (46)$$

The constants C_2 and C_4 can be evaluated from the boundary condition that $\phi=V_1/2$ at $z=\delta$. Since A_z must be an even function of z, $C_2=-C_4$, $C_3=C_1$, and

$$A_z(a,z)=-\frac{j}{c}\left[C_1\cos\beta z+\frac{\frac{V_1}{2}+C_1\sin\beta\delta}{\cos\beta\delta}\sin\beta|z|\right] \qquad (47)$$

If the base gap $\delta\to 0$, this reduces to

$$A_z(a,z)=-\frac{j}{c}\left(C_1\cos\beta z+\frac{V_1}{2}\sin\beta|z|\right) \qquad (48)$$

The vector potential is related to the currents by

$$\mathbf{A}(r,z)=\frac{\mu}{4\pi}\int\frac{\mathbf{J}e^{-j\beta R}}{R}dv \qquad (49)$$

where the integration covers the volume in which there are currents. For the circular array, this gives

$$A_z(r,z)=\frac{\mu}{4\pi}\sum_{i=1}^{m}\int_{-h}^{h}I_{zm}(\bar z)\frac{e^{-j\beta R_{mi}}}{R_{mi}}d\bar z \qquad (50)$$

If we now set $r=a$, we may equate the result to the expression in equation 48, giving

$$\sum_{i=1}^{m}I_{zm}(\bar z)\frac{e^{-j\beta R_{mi}}}{R_{mi}}d\bar z=-j\frac{4\pi}{\eta}\left(C_1\cos\beta z+\frac{V_1}{2}\sin\beta|z|\right) \qquad (51)$$

Each element must have a current distribution $I_{zi}(z)$ which satisfies a similar integral equation, equation 1 of the text.

Appendix II

The integral equation 4 of the text is

$$\int_{-h}^{h}I_z^{(n)}(\bar z)K^{(n)}(z,\bar z)d\bar z=-j\frac{4\pi}{\eta}\left(C_n\cos\beta z+\frac{V^{(n)}}{2}\sin\beta|z|\right) \qquad (52)$$

If we let $I^{(n)}(z)=I^{(n)}f^{(n)}(z)$, we can define a function $g^{(n)}(z,\bar z)$ by

$$g^{(n)}(z,\bar z)=f^{(n)}(\bar z)/f^{(n)}(z) \qquad (53)$$

The function $\psi^{(n)}(z)$ is defined by

$$\psi^{(n)}(z)=\int_{-h}^{h}g^{(n)}(z,\bar z)K^{(n)}(z,\bar z)d\bar z \qquad (54)$$

With these, we may write

$$I^{(n)}(\bar z)=I^{(n)}(z)g^{(n)}(z,\bar z)+I^{(n)}(\bar z)-I_z^{(n)}(z)g^{(n)}(z,\bar z)$$

and the integral equation 52 reduces to

$$I_z^{(n)}(z)\psi^{(n)}(z)+\int_{-h}^{h}[I_z^{(n)}(\bar z)-I_z^{(n)}g^{(n)}(z,\bar z)]K^{(n)}(z,\bar z)\,d\bar z=-j\frac{4\pi}{\eta}\left(C_n\cos\beta z+\frac{V^{(n)}}{2}\times\sin\beta|z|\right) \qquad (55)$$

It can be shown that $\psi^{(n)}(z)$ is proportional to the ratio of the vector potential on the surface of the antenna to the current at that point. This ratio is nearly constant over the length of the antenna. We can then write $\gamma^{(n)}(z)$ as

$$\psi^{(n)}(z)=\psi^{(n)}+\gamma^{(n)}(z) \qquad (56)$$

where $\gamma^{(n)}(z)$ is small. Using this in equation 55 and solving for $I_z^{(n)}(z)$

$$I_z^{(n)}(z)=-\frac{1}{\psi^{(n)}}\left\{\gamma^{(n)}(z)I_z^{(n)}(z)+\int_{-h}^{h}[I_z^{(n)}(\bar z)-I_z^{(n)}(z)g^{(n)}(z,\bar z)]\times K^{(n)}(z,\bar z)d\bar z+j\frac{4\pi}{\eta}\left(C_n\cos\beta z+\frac{V^{(n)}}{2}\sin\beta\bar z\right)\right\} \qquad (57)$$

The integral in 57 is actually zero. Suppose, however, that $g^{(n)}(z, \bar{z})$ is known only as an approximation, in which case it will not be zero, but will be small. The function $\gamma^{(n)}(z)$ is also small. Thus, the only terms on the right that are not small are the trigonometric terms, and these constitute a zeroth-order approximation. The process of iteration described in the text is carried out by substituting this approximation for $I_z(z)$ where it occurs on the right side of equation 57.

Appendix III

The several functions used in expressing the zeroth- and first-order currents and impedances are defined herewith.

The function $\psi^{(n)}(z)$ is

$$\psi^{(n)}(z)=\frac{1}{\sin\beta(h-|z|)}\times \int_{-h}^{h}\sin\beta(h-|\bar{z}|)K^{(n)}(z,\bar{z})d\bar{z} \quad (58)$$

The constant $\psi^{(n)}$ is $|\psi^{(n)}(0)|$. For half-wave dipoles, the expansion parameter $\psi^{(n)}$ becomes

$$\psi^{(n)}=\left|\sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} e^{j\frac{2\pi n i}{m}} C_{dmi}\left(\frac{\lambda}{4},0\right)\right| \quad (59)$$

A function needed later is

$$\gamma^{(n)}(z)=\psi^{(n)}(z)-\psi^{(n)}$$

The function $C_d(h, z)$ is

$$C_d(h,z)=\int_{-h}^{h}\frac{\cos\beta\bar{z}e^{-j\beta[(z-\bar{z})^2+d^2]^{1/2}}}{[(z-\bar{z})^2+d^2]^{1/2}}d\bar{z} \quad (60)$$

This has been tabulated by Mack.[6]

The function $M_0(z)$ is

$$M_0(z)=\sin\beta(h-|z|) \quad (61)$$

and

$$F_0(h)=\cos\beta h \quad (62)$$

The first-order functions $M_1^{(n)}(z)$ and $F_1^{(n)}(h)$ are

$$M_1^{(n)}(z)=F_1^{(n)}(z)\sin\beta h-G_1^{(n)}(z)\cos\beta h+G_1^{(n)}(h)-F_1^{(n)}(h) \quad (63)$$

in which

$$F_1^{(n)}(z)=-F_{0z}(z)\gamma^{(n)}(z)-\int_{-h}^{h}[F_{0z}(\bar{z})-F_{0z}(z)g(z,\bar{z})]K^{(n)}(z,\bar{z})d\bar{z} \quad (64)$$

$$F_1^{(n)}(h)=-\int_{-h}^{h}F_{0z}(\bar{z})K^{(n)}(h,\bar{z})d\bar{z} \quad (65)$$

$$G_1^{(n)}(z)=-G_{0z}(z)\gamma^{(n)}(z)-\int_{-h}^{h}[G_{0z}(\bar{z})-G_{0z}(z)g(z,\bar{z})]\times K^{(n)}(z,\bar{z})d\bar{z} \quad (66)$$

$$G_1^{(n)}(h)=-\int_{-h}^{h}G_{0z}(\bar{z})K^{(n)}(h,\bar{z})d\bar{z} \quad (67)$$

and

$$g(z,\bar{z})=\frac{\sin\beta(h-|\bar{z}|)}{\sin\beta(h-|z|)} \quad (68)$$

By means of algebraic manipulation, all of these can be expressed in terms of the generalized sine and cosine integral functions.

References

1. Uber die Richtcharakteristic von in einer Ebene Angeordneten Strahlern, **H. Stenzel.** *Elektrishe Nachrichten-Technik*, Berlin, Germany, vol. 6, May 1929, pp. 165–181.

2. Bidrog Til Teorien for Antennesystemer med Hel Eller Delvis Rotationssymetri, **H. Lottrup Knudsen.** I Kommission Hos Teknisk Forlag, Copenhagen, Denmark, 1953.

3. Pattern Synthesis for Antenna Arrays on Circular, Elliptical, and Spherical Surfaces, **R. H. DuHamel.** *Technical Report no. 16*, Electrical Engineering Research Laboratory, University of Illinois, Urbana, Ill., May, 1952.

4. Theory of Linear Antennas, **R. King.** Harvard University Press, Cambridge, Mass, 1956, chap. 3.

5. Theoretical Investigations in Transmitting and Receiving Antennae, **Erik Hallen.** *Nova Acta Regiae Societatis Scientiarum Upsaliensis*, Uppsala, Sweden, vol. 11, 1938, p. 185.

6. Tables of the Integrals $C_d(h,z)$ and $S_d(h,z)$, **R. Mack.** Cruft Laboratory, Harvard University, Cambridge, Mass. (To be published.)

7. Antennes a Reyonnement Zenithal Reduit, **H. Chireix.** *L'Onde Electrique*, Paris, France, vol. 15, July 1936, pp. 440–56.

8. Current Distribution and Terminal Impedance of Circular Antenna Array, **C. E. Hickman.** *Master's Thesis*, University of Tennessee, Knoxville, Tenn., Aug. 1960.

9. The Use of a Ring Array as a Skip Range Antenna, **J. D. Tillman, W. T. Patton, C. B. Blakely, F. V. Schultz.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 43, Nov. 1955, pp. 1655–60.

10. An Electronically Scanned Circular Antenna Array, **H. P. Neff, J. D. Tillman.** *Convention Record*, Institute of Radio Engineers, pt. I, 1960, pp. 41–47.

A Proposed Electronic Method for High-Speed Teleprinters

N. N. BISWAS
ASSOCIATE MEMBER AIEE

IT IS a common experience among communication engineers today that much telegraph apparatus which was originally developed for conventional telegraph applications is finding increasing use in other fields, notably in the developing art of data communication. This factor has its inevitable influence on the design of telegraph equipment. In this paper a new method for transmission and reception of teleprinter code is described. This is an improvement over the author's original proposal published in 1958.[1]

Transmission and Reception

Fig. 1(A) shows the 7.5-unit start-stop teleprinter code for the word NEW while Fig. 1(B) shows the proposed new method, assuming 20 ms (milliseconds) as the duration of each code element in both systems to facilitate comparison. Among the noteworthy features of the new method are the following:

1. The start signal is replaced by a reset signal *RS*, consisting of two positive pulses separated by a NO pulse, each of 10-ms duration.

2. The stop pulse is replaced by a negative pulse of 10-ms duration.

3. A space is transmitted by a NO pulse and a mark by a positive pulse of 10-ms duration. Each space or mark is preceded by a negative pulse of the same duration.

4. During periods of no transmission, that is, when the line is idle, there is no current flowing in the line. In the present system, marking current flows into the line during the idle period.

5. The new code always needs a double-current system for transmission, whereas the present code can be transmitted by either a single- or a double-current system.

From Fig. 1(B) it can be seen that each character has 14 units, with each unit of 10-ms duration.

The teleprinter receiver in this method is completely electronic. It recognizes the teleprinter signal serially, but feeds a mechanical printer with a 5-wire parallel input. The block diagram of such a receiver is shown in Fig. 2. The incoming signal encounters a gate and a check circuit. The gate diverts the negative pulses to the counter and the positive

---

Paper 61-29, recommended by the AIEE Telegraph Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 20, 1960; made available for printing November 16, 1960.

N. N. Biswas is with the University of Roorkee, Roorkee, India.

The author wishes to thank W. Y. Lang of the Bell Telephone Laboratories for many valuable discussions about the proposed method described in this paper.

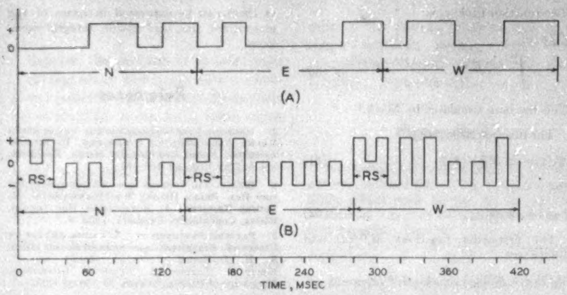


Fig. 1. Current waveform for the word "new" in the existing method (A) and in the proposed method (B)

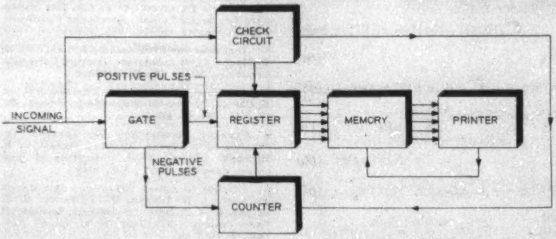


Fig. 2. Block diagram of electronic receiver

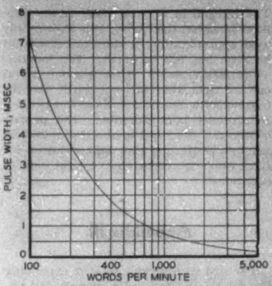


Fig. 3. Pulse width versus speed

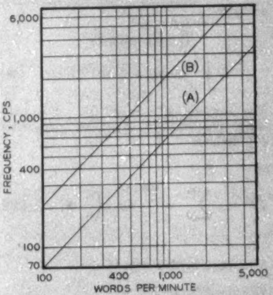


Fig. 4. Signaling frequency versus speed

A—Fundamental frequency
B—Third harmonic

pulses to the register, which can store five bits of information in five locations. The reset signal is unique in that it has two consecutive positive pulses with no negative pulse between. The check circuit produces a pulse output only when it receives the reset signal. This pulse resets the counter and register to the initial conditions. When the counter receives the first negative pulse, the register is conditioned to store the incoming information, which may be a positive pulse in case of mark or no pulse in case of space, in the first location. In this way the register stores the code of the incoming character as a parallel output. When the counter receives the last and sixth negative pulse, the code stored in the register is transferred to the memory. The register then returns to its initial condition and becomes free to receive the next code train. When the next code train is being received by the counter and register, the memory actuates the printer to print the character stored in it.

It can be seen that this method does not require synchronization between the transmitter and receiver. Thus in this system the teleprinter receiver can receive a signal transmitted at any speed provided the pulse is broad enough to be within the operating limit of the electronic elements of the counter and register units. Since an electronic circuit can operate even in a few microseconds (μsec), the speed of transmission can be enhanced to a great extent. The ultimate limit, however, will depend on the type of output required. Assuming six characters per word, speed in words per minute (wpm), S, has the following relation to pulse width in ms, W,

$$W = 5{,}000/(7S) \qquad (1)$$

The curve plotted in Fig. 3 shows that even at a speed as high as 3,000 wpm the pulse width is as broad as 240 μsec. The fundamental frequency of signaling, together with the third harmonic, are plotted against speed in Fig. 4, based on the relation

$$f = 0.7S \qquad (2)$$

where f is the fundamental frequency in cycles per second (cps) and S is in wpm.

Carrier Operation without VFT

For long-distance telegraphic communication, teleprinters generally work with a voice-frequency carrier telegraph terminal (VFT). The VFT in its turn operates on a carrier telephone channel. In this system the teleprinter signal can be made to modulate the telephone channel directly, thereby eliminating the VFT completely. It can be computed from equation 2 or Fig. 4 that with a bandwidth of 3,000 cps a speed of 1,428 wpm can be achieved, assuming that the third harmonic is essential to maintain signal shape. Thus if the teleprinter transmitter and receiver can operate at such a high speed, the VFT can be eliminated by the use of two high-speed teleprinters, one at the transmitting and the other at the receiving end.

While the electronic portion can be easily designed for high speed, the printing mechanism offers some difficulties. However, high-speed printers have been developed to meet the demand of increased speed.[3] If high-speed printers are not available, then the signal can be recorded on high-speed reperforators. The reperforated tape then can be driven slowly to obtain page or tape print-outs on standard equipment.

Conclusions

In the proposed system the teleprinter receiver has greater operating freedom. It no longer remains a single-speed machine, but can receive messages at a continuously variable speed without requiring manual adjustment. Moreover, the accuracy of reception depends on the detection of pulses and their polarities, rather than on their relative durations. Hence bias and other types of distortion which critically affect existing receivers are less likely to introduce error. If, somehow, the line breaks or one negative pulse is lost during transmission, the error is not cumulative, but is confined to one character only as the reset signal preceding the code elements brings the receiver back to the initial condition.

Except for the tape mechanism at the input and the printer or reperforator at the output, the teleprinter is completely electronic and thus can be made of solid-state devices. The design of electronic regenerative repeaters is also simplified as a timing circuit is not necessary. In addition to these advantages, the elimination of the voice-frequency terminal should result in great savings in both primary cost and maintenance.

In closing it should be pointed out that although the proposed method has been described in this paper with particular reference to teleprinters, it can also be extended to the field of digital data communication.

References

1. A Proposed New Method for Transmission and Reception of Teleprinter Code, N. N. Biswas. *Journal*, Institution of Engineers, (India), Calcutta, India, vol. 38, no. 10, pt. 2, June 1958, pp. 977–84.

2. Development of Creed Telegraph Apparatus, B. Brooke-Wavel. *Electrical Communication*, New York, N. Y., vol. 35, no. 3, 1958, pp. 179–84.

A Low-Cost Number Identifier for Small Telephone Systems

T. R. REDINGTON
NONMEMBER AIEE

INABILITY TO determine, accurately and economically, the calling line number in connection with direct distance dialing has hindered the extension of this service to small telephone systems. Manual identification methods, while relatively low in equipment costs, are vulnerable to operator error and subscriber fraud. On the other hand, most currently available automatic number identification equipment cannot be applied economically to systems having a low volume of toll traffic.

The number identification problem is of particular importance to the independent telephone industry which serves many small exchanges. Discussed herein is a new identifier, developed to handle this problem economically. A. A. Jorgensen of the Stromberg-Carlson Company conceived the basic idea, which was developed jointly by Mr. Jorgensen and the author of this paper.

General Approach

Most existing automatic number identification methods were designed around fairly complex common equipment. The equipment individual to each line or terminal, on the other hand, is relatively simple and inexpensive. The concentration of common equipment results in a fixed cost, which is so high that use of direct distance dialing by offices of less than a few thousand terminals is economically unfeasible.

In the design of the identifier under discussion, however, a balance has been maintained between common and individual equipment in order to serve more adequately offices of less than 10,000 terminals. The use of active elements in the individual terminal equipment, while resulting in somewhat higher costs per terminal, makes possible the use of simpler common equipment with a distinct cost advantage in the over-all system.

The basic identifier principle has been applied in two forms. The first is a serial identifier with a single detector and integral outpulser providing serial dial pulse output. This version has a capacity of 300 busy-hour calls and is primarily designed for application to the *XY* toll-ticketing system, which stores data in decimal form on a digital magnetic recorder. The second version of the identifier is equipped with ten detectors, and provides a parallel coded output. The parallel identifier has a basic capacity of 2,400 busy-hour calls. External outpulsers or identifier senders, providing either dial pulse or multifrequency output, are equipped as required by traffic load. A parallel identifier has also been designed with an itegral multifrequency outpulser.

The system of identification under discussion has been developed primarily for application to step-by-step switching systems. It may be applied, however, to virtually any electromechanical switching system employing a third wire, or sleeve lead.

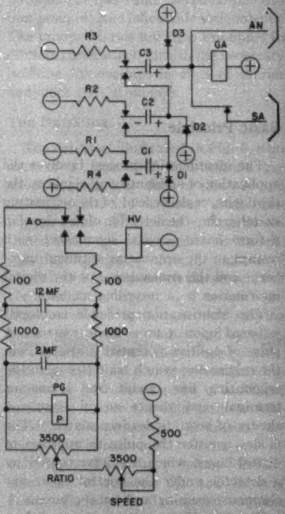


Fig. 1. Pulse generator

AN—To detector anode circuit
SA—To detector starter anode circuit and sleeve pulse-shaping network

---

Paper 61-69, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 31, 1960; made available for printing December 7, 1960.

T. R. Redington is with Stromberg-Carlson Company, Rochester, N. Y.

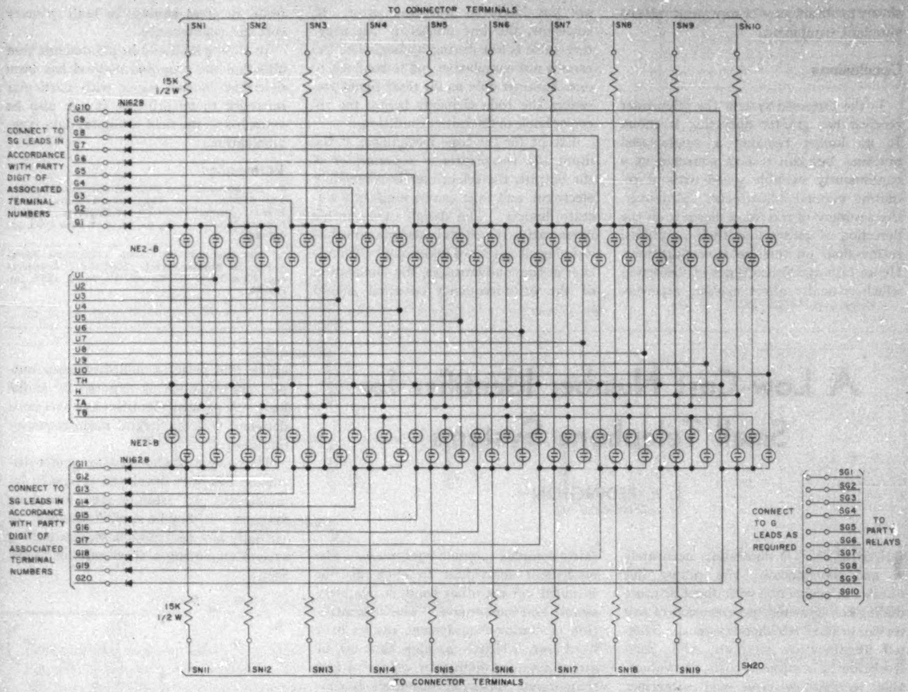

TO CONNECTOR TERMINALS
SN1
SN2
SN3
SN4
SN5
SN6
SN7
SN8
SN9
SN10
15K
1/2 W
IN1628
G10
G9
G8
G7
G6
G5
G4
G3
G2
G1
CONNECT TO SG LEADS IN ACCORDANCE WITH PARTY DIGIT OF ASSOCIATED TERMINAL NUMBERS
NE2-B
U1
U2
U3
U4
U5
U6
U7
U8
U9
U0
TH
H
TA
TB
NE2-B
IN1628
G11
G12
G13
G14
G15
G16
G17
G18
G19
G20
CONNECT TO SG LEADS IN ACCORDANCE WITH PARTY DIGIT OF ASSOCIATED TERMINAL NUMBERS
15K
1/2 W
SN11
SN12
SN13
SN14
SN15
SN16
SN17
SN18
SN19
SN20
TO CONNECTOR TERMINALS
SG1
SG2
SG3
SG4
SG5
SG6
SG7
SG8
SG9
SG10
CONNECT TO G LEADS AS REQUIRED
TO PARTY RELAYS

Fig. 2. Terminal-per-station matrix arrangement

Basic Principle

The identification process involves the application of an identifying signal to the third wire, or sleeve lead, of the originating switch train; the detection of this signal in a form having digital significance with respect to the originating terminal number; and the transmission of the digital information to a recording medium.

The identification principle employed is based upon a neon diode matrix. A pulse of positive potential is applied, via the originating switch train sleeve, to the originating line circuit and connector terminal, and thence to an associated cluster of neon diodes in a matrix. The diodes transfer the pulse to appropriate digital buses, which are interrogated by a detector under the control of a synchronous scanning and pulsing circuit.

The identifier consists, then, of four basic functional blocks: a high-voltage pulse generator, a neon diode matrix, a gas tube detector, and a relay scanning and control circuit.

The High-Voltage Pulse Generator

Two mercury-wetted contact relays, one of which is bistable and polarized, comprise the activating means for generating high-voltage pulses from the office battery supply. The pulses provide the anode potential for the detector as well as the identification signal to the sleeve lead.

Fig. 1 shows the pulse generator; for descriptive simplicity, contact-protection networks have been omitted from the drawing. In Fig. 1, the polar relay is arranged in a self-interrupting, capacitor-timed configuration, which is activated by the application of positive potential to point *A*. Positive pulses, at the rate of 10 pulses per second and a ratio of 38% make, are delivered to relay *HV* which follows the pulses.

Relay *HV* provides polarity switching for a voltage tripler circuit, consisting of capacitors *C1*, *C2*, and *C3;* resistors *R1*, *R2*, *R3*, and *R4;* and silicon diodes *D1*, *D2*, and *D3*. When relay *HV* is normal, the capacitors charge to office battery potential. Relay *HV*, in operating, (1) reverses the potential at the negative terminal of *C1;* (2) connects capacitors *C1*, *C2*, and *C3*, in series; and (3) cuts off diodes *D1*, *D2*, and *D3*, which causes an output voltage equal to the sum of the charged voltages of *C1*, *C2*, and *C3* to appear at the positive terminal of *C3*. This potential is fed directly to the main anode circuit of the cold cathode tube in the detector circuit. Relay *GA* is a fast-acting relay, which operates 1 ms (millisecond) after the appearance of the positive potential at the junction of *C3* and *D3*. Positive potential is gated through break contacts of relay *GA* to the starter anode circuit of the detector and to a pulse-shaping network for application to the sleeve. Consequently, when the pulse generator is operating, positive pulses of 38-ms duration at the base line are delivered to the anode circuit of the detector at the rate of 10 pulses per second, while pulses of 1-ms duration are delivered to the starter anode circuit of the detector and the sleeve pulse-shaping network. The sleeve pulse is shaped, prior to insertion in the sleeve circuit, to round off the steep front, to minimize crosstalk, and to integrate steps in the rising front caused by unfavorable contact stagger in relay *HV*.

The slope of the exponential decay of the pulse, in the 1-ms period prior to the operation of relay *GA*, is a function of the office sleeve load. Since the pulse, after its application to the sleeve, must provide the energy to hold the switch-through relays of the selectors in the originating switch train, its decay characteristic is dependent on the number of selectors in the train. The generator is rated at 150 ohms, maximum load.

The relays used in the pulse generator are of the mercury-wetted contact type because of their inherent reliability and bounce-free characteristic, both necessary to the application described.

THE MATRIX

The matrix consists of a cluster of four neon diodes and a resistor for each connector terminal. A selenium diode for party-gating purposes is added to the clusters serving those connector terminals assigned to multiparty lines in terminal-per-station groups. Clusters consisting of three neon diodes and a resistor are used with those connector terminals assigned to multiparty lines in terminal-per-line groups. The anodes of all diodes in a cluster are connected in parallel through the resistor to the sleeve lead of the associated connector terminal. The cathodes of the diodes are each connected to one of ten, thousands, hundreds, tens, or units buses in accordance with the corresponding digits of the connector terminal number. Fig. 2 is a schematic drawing of the arrangement of the matrix components for terminal-per-station multiparty service, while Fig. 3 illustrates the physical arrangement of the corresponding printed circuit card.

Neon diodes of the *NE2* type exhibit characteristics for use as a gating element in an identification matrix. The d-c resistance drops from greater than 500 megohms in the deionized or nonconducting state to about 8,000 ohms when ionized. For all practical purposes, this characteristic completely eliminates the marginal signals on matrix buses which are characteristic of resistor matrices. Consequently, signal detection becomes a nonmarginal go or no-go process, readily accomplished by simple detection circuitry.

In operation, the matrix transfers the identification signal appearing at the sleeve lead of a connector terminal to 4 of a possible 40 multiple buses having digital significance with respect to the terminal number. Since the anodes of the neon diodes of the matrix cluster assigned to each connector terminal are connected to its sleeve lead through a limiting resistor, a voltage appearing on the sleeve also appears on the anodes of the diodes. The identification signal at the connector terminal sleeve is a positive pulse of 120 volts nominal amplitude. This amplitude is sufficient to force into breakdown a diode whose cathode is suitably terminated; this, in turn, causes a positive voltage to appear on the multiple bus to which the diode is connected.

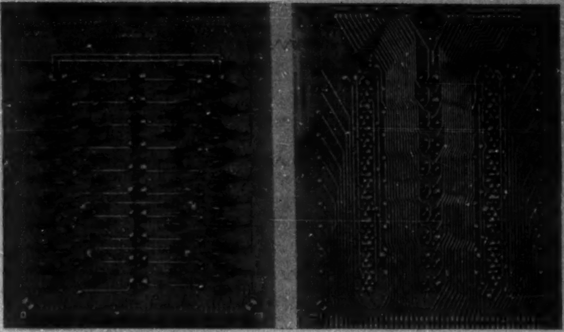


Fig. 3. Terminal-per-station matrix card

Multiparty lines of the terminal-per-line type present no particular problem. Since each station on a line is distinguished only by the units digit of its number, it shares common thousands, hundreds, and tens digits with the other parties on the line. Consequently, matrix clusters are assigned on a per-line basis to determine the first three digits of the number only. The units digit is determined from the party number, or circle digit, dialed by the originating party and transmitted directly to the recording equipment.

In the case of terminal-per-station lines, however, the process becomes somewhat more complex. The party number is dialed by the originating party and is used to activate a gating process. Fig. 4 illustrates the party-gating principle. Normal contacts of relays, assigned to each party, maintain the cathodes of the selenium diodes of the matrix clusters at ground potential, thus shunting to ground the identification signal appearing at the clusters of all of the parties on the line except the one whose party digit has been dialed. Where 2-party service is involved, the party relays are activated by party-detection circuitry rather than a subscriber-dialed digit.

In addition to its function in determining the terminal number that is originating a particular call, the matrix also serves to determine the class of multiparty service and the office code, should more than one code be assigned within one 10,000-terminal group. Each of these functions is accomplished by grouping thousands buses in accordance with terminal-per-line and terminal-per-station grouping and office code assignments. The grouped buses are then examined in discrete steps, which furnish the necessary isolation for overlaps in class of service and office code groupings.

THE DETECTOR

The detector, illustrated in Fig. 5, consists of a type *5823* cold-cathode gas triode furnished in dual, a relay in the anode circuit, and a Zener diode biasing network. The anode voltage is obtained from the pulse generator circuit, as is the bias supply. It is the function of the detector to respond to the presence of the identifying signal on a matrix bus. Because of the instantaneously varying nature of the sleeve load caused by selectors searching over a sleeve being identified, the signal appearing on a matrix bus cannot be guaranteed to be of sufficient amplitude to fire the cold-cathode tube directly. Consequently, a biasing arrangement is necessary. A bias of +60 volts is developed by a voltage divider, consisting of resistors *R1* and *R2* and Zener diode *Z1*. To guard against spurious signals, the tube is biased only during the 1 ms of each pulse generator cycle that a legitimate signal can possibly appear.

The detector circuit is equipped with two cold-cathode tubes to enhance its

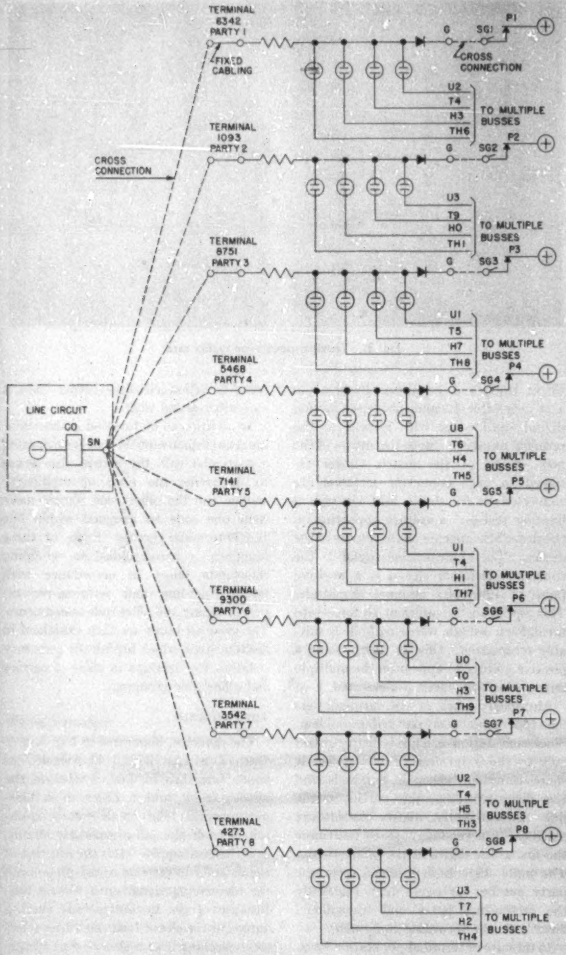

TERMINAL 6342 PARTY 1
FIXED CABLING
G
SG1
P1
CROSS CONNECTION
U2
T4
H3
TH6
TO MULTIPLE BUSSES
TERMINAL 1093 PARTY 2
CROSS CONNECTION
G
SG2
P2
U3
T9
H0
TH1
TO MULTIPLE BUSSES
TERMINAL 8751 PARTY 3
G
SG3
P3
U1
T5
H7
TH8
TO MULTIPLE BUSSES
TERMINAL 5468 PARTY 4
G
SG4
P4
LINE CIRCUIT
CO
SN
U8
T6
H4
TH5
TO MULTIPLE BUSSES
TERMINAL 7141 PARTY 5
G
SG5
P5
U1
T4
H1
TH7
TO MULTIPLE BUSSES
TERMINAL 9300 PARTY 6
G
SG6
P6
U0
T0
H3
TH9
TO MULTIPLE BUSSES
TERMINAL 3542 PARTY 7
G
SG7
P7
U2
T4
H5
TH3
TO MULTIPLE BUSSES
TERMINAL 4273 PARTY 8
G
SG8
P8
U3
T7
H2
TH4
TO MULTIPLE BUSSES

Fig. 4. Terminal-per-station gating arrangement

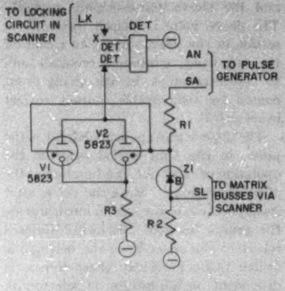

TO LOCKING CIRCUIT IN SCANNER
LX
X
DET
DET
DET
AN
SA
TO PULSE GENERATOR
R1
V1 5823
V2 5823
Z1
B
SL
TO MATRIX BUSSES VIA SCANNER
R3
R2

Fig. 5. Detector circuit

over-all reliability. Since the first tube to fire instantly starves out the other, only one tube functions at any given detection, the other tube serving as a stand-by.

When a matrix bus carrying the identifying signal is connected to the detector, the positive voltage appearing on the bus becomes additive with the bias on the starter anode circuit. This results in a potential difference between starter anode and cathode sufficient to fire one of the tubes, and operate the DET (detector) relay, which locks independently to a preliminary make-contact. The DET relay then performs the necessary switching functions to program the sequence of the scanning and control circuit.

The Scanning and Control Circuit

This circuit uses conventional relay techniques to control the synchronous scanning and pulsing action on which the entire identification process depends. It also serves to associate the identifier with a particular call. Different scanning and control configurations are used with the serial-and parallel-output identifiers. The same program and matrix-reduction modules are used in both, since the function is essentially the same regardless of final output.

The program module serves to control the sequence of the identification process. That is, it controls the necessary steps (1) to determine the class of multiparty service, (2) to determine and transmit the office code, and (3) to determine and transmit each of the digits of the terminal number.

The matrix reduction module, responding to the command of the program module, serves to connect sequentially each group of ten matrix buses to a detection multiple.

In the serial identifier, the detection multiple is scanned by a single detector, which is connected sequentially to each of the ten leads of the multiple by a relay-counting circuit. As each multiple lead is scanned, the identifier output lead is impulsed, resulting in the transmission of a dial pulse to the recording equipment. When a detection is made, impulsing ceases for a period of 600 ms, after which scanning and impulsing begin again. The 600-ms pause serves to mark the end of a digit in the recording equipment, and corresponds to the interdigital interval of conventional dialing. The transmission of each digit comprising the terminal number is accomplished by scanning each of the four groups of matrix buses, via the detection multiple, with the aid of the program- and matrix-reduction modules.

The parallel-output identifier is equipped with ten detectors, which are connected simultaneously to each of the four groups of matrix buses via the detection multiple. Each cycle of the pulse generator results in the detection of a digit, which is transmitted to the registers of an outpulser in coded form.

The scanner and control circuit also functions to associate the identifier with the accessing circuits, via a conventional relay-preference chain. It also alternates calls between dual identifiers and diverts all traffic to one or the other identifier in the event of failure.

Maintenance Aspects

Minimal maintenance requirements and operational simplicity were two objectives of the design of the identifier. The nature of the service for which the identifier is intended precluded reliance on daily routine testing. The serial identifier in particular will often be installed in unattended offices which are visited at less than weekly intervals. The maintenance personnel economically available to the small telephone system must of necessity be capable of handling both inside and outside plant. The average combination man who fulfills this requirement cannot reasonably be expected to possess a depth of electronic knowledge or training.

Every effort consistent with economy and simplicity has been made to insure reliable operation. All marginal relay operations have been avoided and relay spring loads have been held below the usual levels to minimize wear. Since the basic principle of the identifier is dependent on voltage-sensitive devices, the design voltage range was set at 40 to 60 volts, although the normal operational guaranteed range is 44 to 55 volts.

A portable identifier test set has been designed to provide a simple check of the identifier output on known terminal numbers and to record error on a digital display. The combined use of the test set and visual trouble-shooting techniques bring the maintenance requirements of the identifier within the grasp of personnel having some knowledge of electromechanical switching functions. Since there is a visual indication of the firing of diodes in the matrix and the detector tubes, the progress of an identification may be readily followed in the process of fault location. In addition to the test set, only commonly available tools and a neon test lamp are required to maintain this equipment.

Conclusions

The identifier discussed in this paper will assist the independent telephone industry in extending direct distance dialing on an *ANI* basis to exchanges once considered too small for such service. The combination of simple electronic techniques and conventional electromechanical functions results in a simple, economical, and readily maintained system of automatic number identification.

A Completely Transistorized 600-Channel Multiplex System

D. W. SPARKS
NONMEMBER AIEE

THE PAST 10 YEARS have brought forth few improvements in long-haul telephone carrier systems, despite steadily increasing consumer demands. Existing carrier systems were designed primarily for the transmission of speech and have been used mainly for that and for telegraphic transmissions. The common carriers, independents, and private and government-owned systems have done a remarkable job in providing new types of service such as handling of data, Teletype, and facsimile. However, costs have run high because of the need to derate capacities of systems designed primaily for voice transmission or to add suitable devices for correcting delay distortion, amplitude distortion, frequency stability, or signal-noise ratio.

In 1958, the U. S. Air Force saw the need and took the leadership in writing specifications for a long-haul transistorized 600-channel system, *AN-FCC17*, which was intended to set the standard for the 1960's.

Evolution of TCS-600

A system which would meet rapidly changing conditions and requirements was interpreted by General Electric Company to mean a network capable of short, medium, and long haul and low-, medium-, and high-channel capacity. This dictated adopting the principle of infinite building block technique, both electrically and mechanically. All known requirements had to be catered to, as well as important unknown future requirements.

The equipment would require as many channels in as small a space as component and cost development would allow. It was decided that 120 channels, housed into a standard 8-foot rack, would optimize space and cost. With this decision came the problem of wiring. Because of limited space, a building block wiring technique was employed, which allowed many completely different types of units to be plugged into the same shelf spaces.

Producing a system to meet present-day electrical and mechanical demands only was not considered sufficient; other established aims dealt with reliability, ease of maintenance, human engineering, and pleasing yet functional appearance.

Fig. 1 shows a 7-foot relay rack of General Electric Company's *TCS-600* carrier system, equipped with 60 telephone channel modems, or modulator-demodulator circuits. An 8-foot relay rack will accommodate 120 channel modems. Fig. 2 shows the common equipment rack which is expandable in the same rack to accommodate 600 telephone channels.

Brief Over-all Specifications

The *TCS-600* is a completely transistorized (no tubes or relays) frequency division, single side-band, suppressed carrier system. The modulation plan is International Telegraph and Telephone Consultative Committee (CCITT) Group Plan *B* with plug-in carrier and modem units to meet Western Electric Company alloca-

---

Paper 61-111, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 31, 1960; made available for printing December 21, 1960.

D. W. Sparks is with General Electric Company, Lynchburg, Va.

The General Electric Company and the author wish to express gratitude to innumerable company personnel of the Communication Products Department who made possible the system described in the paper. Dr. R. A. Stasior of the Semiconductor Products Department, for his work and advice in developing the transistor modulator, is especially commended as is the engineering staff of the Western Union Telegraph Company whose guidance, advice, and criticism were invaluable.

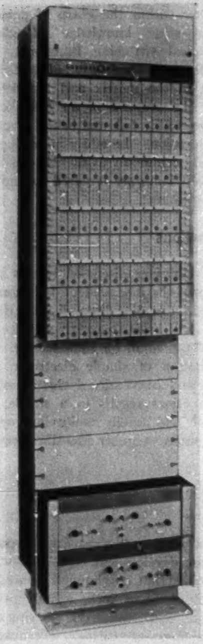


Fig. 1. TCS-600 channel modem rack, equipped for 60 channels

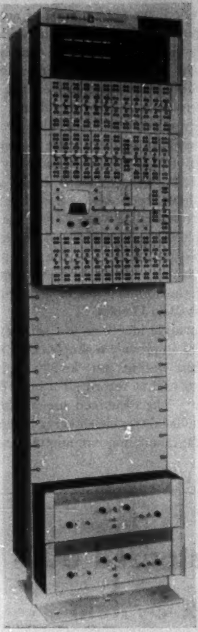


Fig. 2. TCS-600 common equipment rack

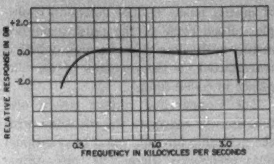


Fig. 3. Typical channel response

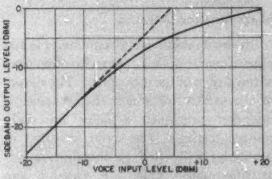


Fig. 4. Limiting characteristics, channel modulator

tions where these differ from CCITT. The channel response, as shown in Fig. 3, for one link is ±1 db (decibel) from 300 cps (cycles per second) to 3,600 cps. The signaling is either 3,825 cps out-of-band or 2,600 cps in-band. Terminations of 600- or 900-ohm, *E* and *M* (*E* refers to incoming or receiving leads and *M* to outgoing or transmittal leads), or 20-cps ringdown may be employed. Ringdown is the term applied to a circuit which is not equipped for dial or automatic operation. The *E* and *M* impulse rate is suitable for direct distance dialing. The modulator is discussed more fully in later paragraphs and illustrated in Fig. 4.

The idle channel noise for a back-to-back terminal is 10 dba (decibels adjusted), referred to a 0-level test tone, and loaded channel noise (CCITT loading) is 20 dba.

Design Requirements

PASSIVE NETWORKS

One of the greatest problems facing the designer of carrier equipment today is delay distortion. In a frequency division system, passive networks play a predominant role and with them come built-in undesirable delay characteristics. Networks which exhibit flat delay are impractical when one is considering 4-kc channels with 700-cps guard bands in a "small space" package. Future carrier systems, even the frequency division type, will not contain these passive networks. At present, the problem is still very real. Filter networks without inherently maximal flat delay characteristics have an attenuation versus phase relationship which has produced the axiom that the flatter the attenuation versus frequency curve in-band, the more constant is the delay. By making a filter network of this type with the widest band possible, a flat delay characteristic is achieved over it. Having achieved a wide band, a small amount of amplitude equalization can be

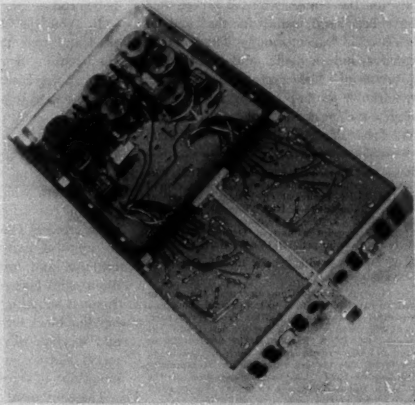


Fig. 5. Channel carrier selector showing plug-in stand-by amplifiers

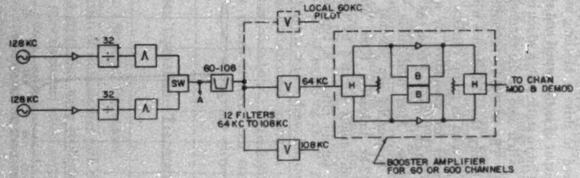

128KC
128KC
32
32
SW
A
60-108
LOCAL 60KC PILOT
12 FILTERS 64KC TO 108KC
64KC
108KC
H
B
B
H
TO CHAN MOD & DEMOD
BOOSTER AMPLIFIER FOR 60 OR 600 CHANNELS

Fig. 6 (left). Master oscillator and channel carrier supply

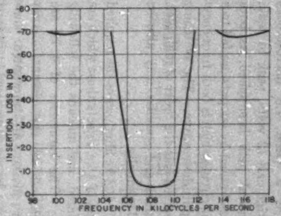


Fig. 7 (right). Typical channel carrier selector filter

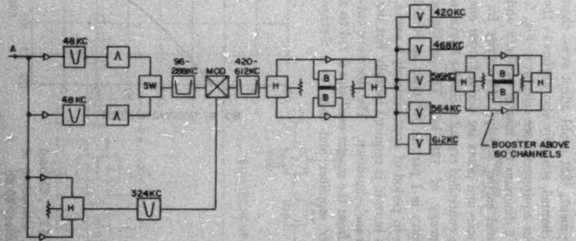

A
48KC
48KC
SW
96-288KC
MOD
420-612KC
324KC
H
B
B
H
420KC
468KC
516KC
564KC
612KC
BOOSTER ABOVE 60 CHANNELS

Fig. 8 (left). Group carrier supply

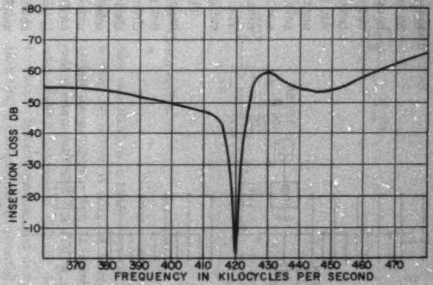


Fig. 9 (right). Typical group carrier selector filter

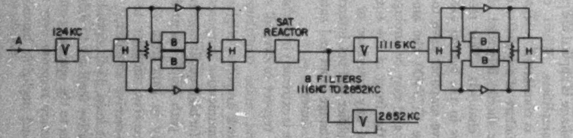

A
124KC
H
B
B
H
SAT REACTOR
8 FILTERS 1116KC TO 2852KC
1116KC
2852KC

Fig. 10 (left). Supergroup carrier supply

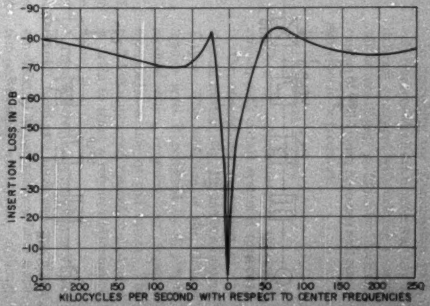


Fig. 11 (right). Supergroup carrier selector filters

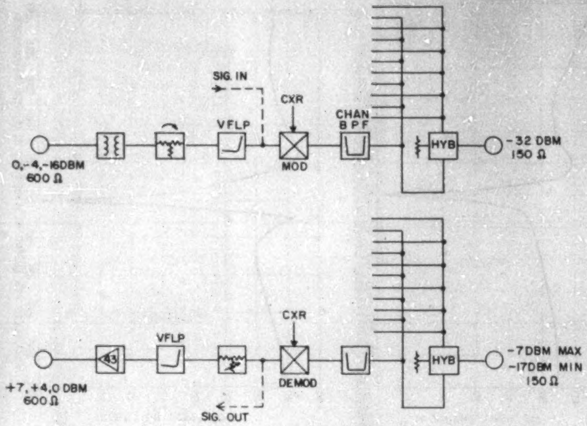


Fig. 12. Channel modem

applied to a channel modem to improve both amplitude and delay. If still flatter delay is required, then active delay equalization can be applied.

Active Networks

Present and future requirements dictate reliability in new systems. The only active elements in the transmission path to which direct voltage is applied are contained in the amplifiers. When cost versus reliability is considered, a reasonable rule is that failure of any active element or network should interrupt only one out of 600 channels. This means that all amplifiers in the system except the channel-receive amplifier have to be stood-by. In the *TCS-600*, the amplifiers are stood-by and may be unplugged separately from the front of the equipment. Failure of any transistor or removal of one amplifier will not interrupt transmission, or change input and output voltages and impedances.

Amplifier design is extremely important, particularly with respect to over-all system stability and noise contribution. The bridge feedback amplifier, originally designed for communication purposes, is an obvious choice for high-quality systems. The transmission path amplifiers used in the *TCS-600* vary in gain from 43 db for the channel-receive amplifier to 35 db for the group-receive amplifier. All amplifiers have at least 65-db open-loop gain. Using negative feedback to reduce this open-loop gain to the operating value gives a flat frequency response over the operating range and negligible harmonic distortion.

The stand-by of active networks also is applied to carrier and power supplies.

Equipment Operation—Carrier Supplies

Master Oscillator and Channel Carrier Supply

The philosophy of having all active equipment, that could affect more than one channel, stood-by also applies to the carrier equipment; see selector in Fig. 5. Fig. 6 diagrams the master oscillator and channel carrier supply. All carrier frequencies and pilots are derived from one oven-controlled crystal oscillator having a frequency of 128 kc. The oscillator stability is 1 part in 10^7 after an initial burn-in period of approximately 3 weeks. This burn-in is required initially but not after each shutdown. The sine wave output from the master oscillator is fed to a buffer amplifier and then to a 5-stage binary divider which divides the 128 kc by 32 to produce 4 kc. This square wave is applied to a blocking oscillator whose output is a pulse 4.8 microseconds (μsec) wide and 12 volts, peak to peak. An identical path is supplied, as both normal and stand-by outputs are fed to a switching circuit which is both frequency- and amplitude-sensitive. The switching time involved is in the order of milli-μsec. The normal and standby may also be switched manually.

The 60–108-kc filter passes the required harmonics of 4 kc, rejects those which are not required, and unloads the following circuitry. The harmonics of 4 kc are then selected by the channel carrier selection filters. All unwanted harmonics of 4 kc are suppressed by at least 68 db; see Fig. 7. The particular carrier is now amplified either by a 60- or 600-channel carrier booster. These booster amplifiers are stood-by and may be unplugged from the front of the equipment as previously mentioned. The booster amplifier is wideband, 60–108 kc, and may be used for any channel frequency booster.

Group Carrier Supply

Fig. 8 shows the group carrier supply. Point *A* marks the pulse output from the 4-kc blocking oscillator. The twelfth harmonic of 4 kc is amplified, selected, and used to drive a blocking oscillator. The output of both normal and stand-by are fed into a static switch which is similar to the switch used in the channel carrier supply. The 96–288-kc unloading filter accepts only those harmonics which will finally be used as group carriers. These frequencies are modulated by a 324-kc carrier which is derived from the original 4-kc pulse. After modulation, the required frequencies are selected by the 420–612-kc unloading filter, amplified, and individually selected by the group carrier supply filters. A typical group

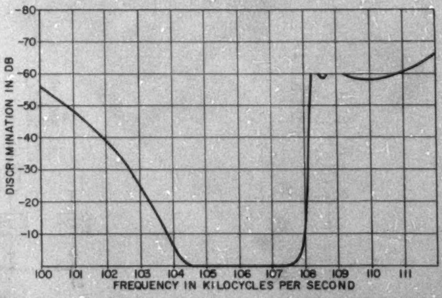


Fig. 13. Channel filter, attenuation at poles>80 db

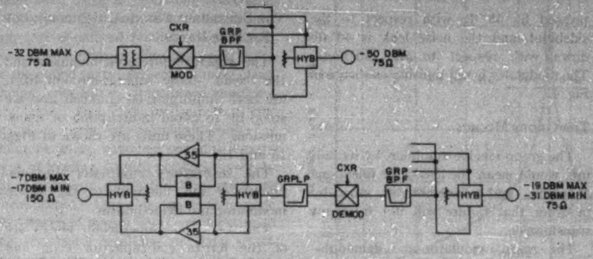


Fig. 14. Group modem

carrier filter is shown in Fig. 9. This filter uses 2 crystals and 2 ferrite-type coils. The carrier supply at this point is capable of operating a 60-channel system. By adding a booster amplifier, the capability is increased to a 600-channel system. Once again the amplifiers are wide-band, 420 kc–612 kc, and may be used for any group carrier frequency. In a CCITT system of this kind, the highest group and the first supergroup use the same carrier frequency and so 612 kc is derived in the group carrier supply for both group and supergroup use. The second supergroup, 312 kc–552 kc, is the basic supergroup and does not require further modulation to place it at line frequencies. A 120-channel system, therefore, does not require the use of a supergroup carrier supply.

Supergroup Carrier Supply

Point *A* in Fig. 10 is the pulse output of the 4-kc blocking oscillator. The thirty-first harmonic is selected and amplified, then used to drive a saturable reactor network, which produces odd harmonics of 124 kc. These frequencies are individually selected, and all other unwanted frequencies are suppressed by at least 75 db, as shown in Fig. 11. The carrier frequencies are then amplified and, at this point, the carrier supply has the capability of supplying carrier power to 8 modulators and 8 demodulators required for 600-channel operation. The amplifier is wide-band, 1,116–2,852 kc. Only one design is used, and may be plugged into any supergroup carrier frequency.

Equipment Operation—Transmission Path

Easy and Difficult Problems

The design of transmission path equipment is a great deal more sophisticated than the pure electronic job of supplying carriers. Filters, amplifiers, modulators, and transformers do not have a simple or easy classification in today's system. If a suitable device is easy in terms of system requirement, then it must be reduced in mechanical size and/or cost, whereupon it becomes a difficult design problem. System analysis and design is such that a computer is an essential tool if the system is not to be outdated before it is manufactured. Human endurance and initiative must also be taken into account when twelfth-degree polynominals are to be solved with 15-digit accuracy.

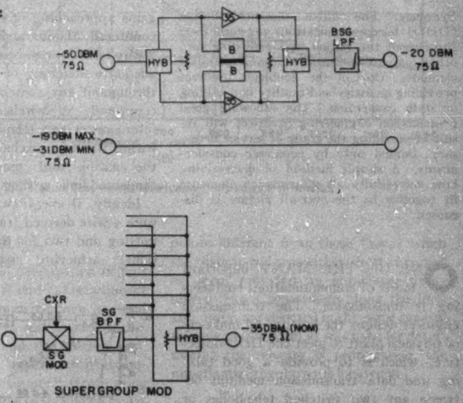
In describing the transmission path, only features which are in some way new or different compared with existing systems will be mentioned; no block-by-block discussion of well-known processes and essentials should be necessary.

The Channel Modem

The only potentiometer in the transmission path of this system is contained in the channel-receive portion of the channel modem, Fig. 12, and is a bridged *T* type. The variation of the channel send-in level is achieved by a plug-in module which contains four separate attenuation-section or pad values. The module is unplugged, rotated, and plugged in again to select a required attenuation value.

The voice-frequency low-pass or VFLP-channel band-pass filter and modulator are designed as one package from a system-performance standpoint, although each is replaceable individually. Channel modem filters are one of the more difficult designs to achieve if consideration is given to performance, cost, and size. An out-of-band discrimination requirement, set 6 db too high, can cost tens of dollars per

Fig. 15 (right). Basic supergroup unit



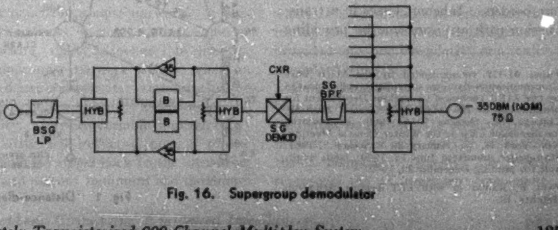


Fig. 16. Supergroup demodulator

channel. The channel filter characteristic is shown in Fig. 13. The filter has only two poles in the finite frequency range, both on the high-frequency side. Poles are also located at zero and infinity of the fifth and third order, respectively. The insertion loss is 2 db and the ripple 0.04 db total, with a reflection of 10%.

The channel response of a back-to-back system is shown in Fig. 3. The limits are ±1 db from 300 cps to 3,600 cps with respect to 1,000 cps.

The channel modulator and demodulator is a 2-transistor balanced modulator and demodulator. The carrier is suppressed by 55 db with respect to the sideband, and the noise leak is 40 db down with respect to the sideband. The modulator is self-limiting as shown in Fig. 4.

The Group Modem

The group-receive amplifier, by its failure, would mean the loss of 12 telephone channels, and therefore is stood-by in order that failure will not interrupt transmission.

The group modulator and demodulator, Fig. 14, is a 4-transistor ring modulator, using identical transistors.

The Supergroup Send and Supergroup Receive Units

The supergroup-receive amplifier and the "send" basic supergroup amplifier, are both amplifying 60 channels and are stood-by to avoid interruption of transmission. These units are shown in Figs. 15 and 16.

The supergroup modulator and demodulator are identical to the group modulator and demodulator.

The first four supergroup filters are of the ferrite coil-capacitor type and the upper six are of the ferrite coil-crystal type.

Telephone Circuit and Equipment Impedances as They Affect Transmission Performance

PAUL F. RADUE
ASSOCIATE MEMBER AIEE

Synopsis: The direct distance dialing (DDD) telecommunications network used throughout the nation and extending into other nations consists of many complex elements. One of the problems involves providing quantity and quality in a talking or data connection. The following good transmission engineering practices will result in providing the grade of service necessary, limited only by economic considerations. A simple method of determining how successfully all transmission elements fit together in the over-all picture is discussed.

ONE OF THE MANY important facets of communications engineering is transmission. The transmission engineer designs the layout and make-up of physical plant to meet the DDD objective, which is to provide a good talking and data transmission medium between any two switched telephones or locations within our nation and beyond our borders. Inherent losses in a transmission path are overcome by providing gains approaching or equal to losses encountered. Gains approaching losses on 2-wire facilities can be realized only when adequate matching of impedances exists throughout any switched connection encountered. A switched connection may interconnect two telephones or locations within the same exchange area or across the nation and may involve many tandemed links as illustrated by Fig. 1.

Ideally, if every switched connection were 4-wire derived (two wires for transmitting and two for the receiving directions), achieving essentially zero db (decibel) net loss, transmission media regardless of length would be relatively easy. Unfortunately, economics, technique, and past practices have dictated extensive use of 2-wire switched offices in nearly every case illustrated by Fig. 1. Since carrier-derived facilities are basically 4-wire in nature, hybrids are required at every 2-wire switched location. By the same token, physically derived facilities are basically 2-wire and would require hybrids in order to interconnect on a 4-wire basis. Until such time as 4-wire switching becomes standard, it will be necessary to improve upon impedance matches in 2-wire switched offices to approach that of a 4-wire switched office.

This is an effort aimed toward a better understanding of fundamental transmission problems encountered in the telecommunications field, and of methods available to achieve the desired results.

Basic Design Objectives

Each individual facility, beginning with the subscriber's telephone set, connecting

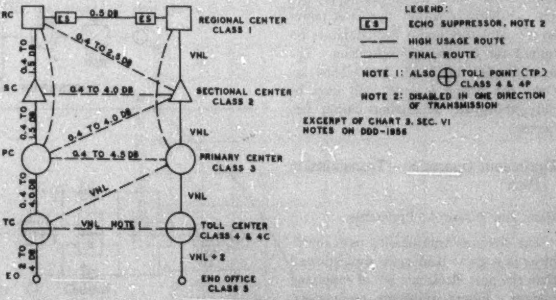


Fig. 1. Distance-dialing range, via net loss (VNL)

Paper 61-117, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960; re-presented for discussion only at the Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted June 13, 1960; made available for printing December 29, 1960.

Paul F. Radue is with ITT-Kellogg, Arlington Heights, Ill.

cable pair, local or interoffice switching, toll-connecting trunk, toll center switching, carrier-derived facility to each succeeding higher-class office and back down to the called party, presents, among other things, numerous impedance-matching problems. Problems in matching impedances occur at switching points as different gages, capacity per unit length, type of loading, hybrid networks, switching equipment characteristics, etc., are encountered. Our objective is to provide the degree of match necessary for attaining satisfactory talking or data transmission throughout any possible tandemed connection.

Maximum transfer of energy between two facilities occurs when conjugate impedances exist. Therefore, should the source impedance be $R+jX$, the load impedance will need to be $R-jX$. The series-connected conjugate impedance results in cancellation of the reactive j components via a series-resonant condition. Parallel junction of conjugate impedances also results in cancellation of reactive components and realization of an antiresonant condition. So, conjugate $R+jX$ and $R-jX$ resultant is:

$$\frac{R^2+X^2}{2R}$$

Conjugate impedance matching is not insisted upon on communication facilities as it may incur an impedance irregularity. Since relatively minute power levels are used, no appreciable amount of power is lost due to matching image impedances in lieu of conjugate impedances. It so happens that at switching points image impedance matching generally meets conjugate impedance-matching requirements also. Image impedance occurs when $R+jX$ of the source is equal to $R+jX$ of the load or $R-jX=R-jX$.

Since impedance matching is our theme, it might be well to discuss how characteristic impedance of a line may be determined by making an impedance measurement with the distant end open and then short-circuited. The characteristic impedance will be

$$Z=\sqrt{Z_0 Z_s}$$

at any particular frequency where Z_0 equals the open-circuit impedance and Z_s equals the short-circuit impedance.

Considerable confusion seems to exist in the field as to the method used in locating an impedance irregularity. From the average difference between minimum or maximum peaks seen on a distant-end short-circuited or open-impedance curve, the speed of a signal traveling along a facility may be calculated as follows:

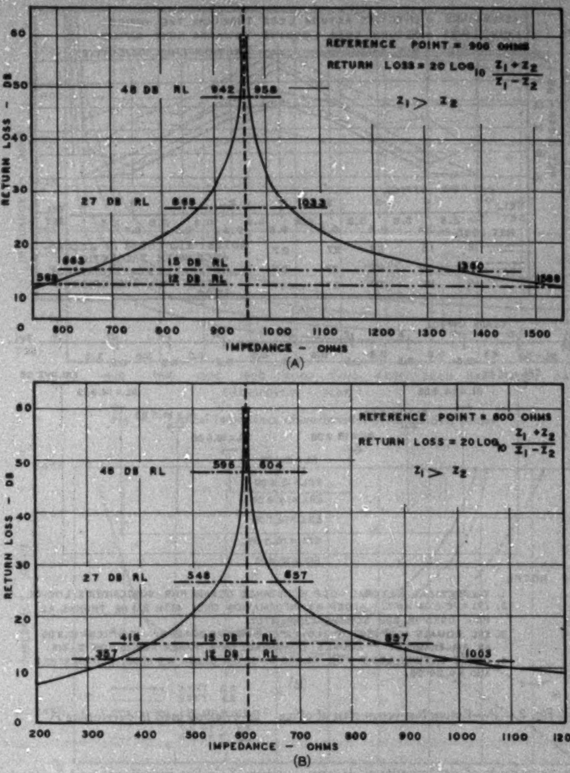


Fig. 2. Return loss versus termination impedance mismatch

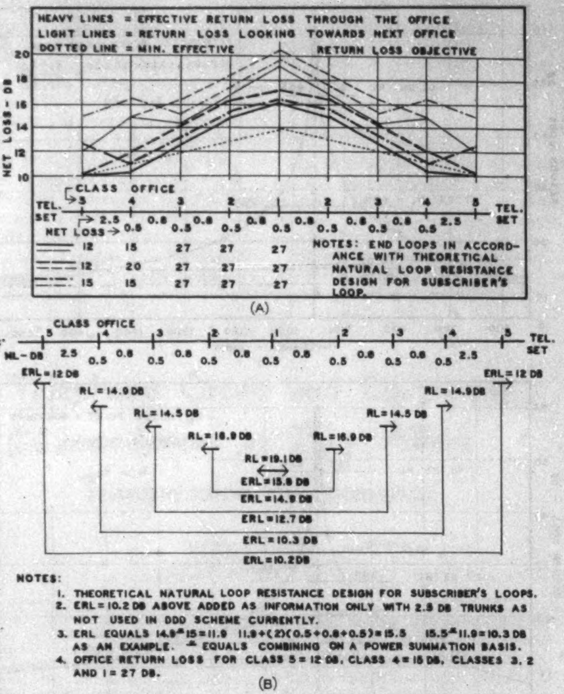
$$V=2(f_2-f_1)d \qquad f_2>f_1$$

where V = velocity of propagation in mps (miles per second), f_2 and f_1 are adjacent maximum or minimum frequencies, and d = actual distance in miles existing between testing end and the distant end. The location of an impedance irregularity is determined by properly terminating the distant end of a line in its known image impedance with the use of a precision network or a compromise network. A precision network is used when a cable or open-wire facility is looked into directly; a compromise network is used when an impedance compensator is used on a loaded cable or on the hybrid of a carrier-derived facility. Any appreciable impedance irregularity will reflect a portion of the transmitted signal back to the measuring end and present minimum and maximum points different from those "seen" when the distant end was either short-circuited or open, provided an impedance irregularity exists somewhere along the line. If f_2 and f_1 are selected in the same general frequency spectrum as was used for determining the velocity of propagation, the impedance irregularity is located at

$$D=\frac{V}{2(f_2-f_1)} \qquad f_2>f_1$$

where D is the distance in miles from the testing end to an impedance irregularity, provided only one irregularity or a major irregularity exists.

Detailed impedance measurements provide resistance and reactance values at a particular frequency. In rectangular coordinates, the X-axis usually represents the R component, whereas the Y-axis represents the jX reactance component.



NOTES:

1. THEORETICAL NATURAL LOOP RESISTANCE DESIGN FOR SUBSCRIBER'S LOOPS.
2. ERL = 10.2 DB ABOVE ADDED AS INFORMATION ONLY WITH 2.5 DB TRUNKS AS NOT USED IN DDD SCHEME CURRENTLY.
3. ERL EQUALS 14.9≛15 = 11.9 11.9 + (2)(0.5 + 0.8 + 0.5) = 15.5 15.5≛11.9 = 10.3 DB AS AN EXAMPLE. ≛ EQUALS COMBINING ON A POWER SUMMATION BASIS.
4. OFFICE RETURN LOSS FOR CLASS 5 = 12 DB, CLASS 4 = 15 DB, CLASSES 3, 2 AND 1 = 27 DB.

(B)

Fig. 3. A—Return loss versus class of office. B—Method used in developing A

The magnitude of an impedance at any particular frequency will then be equal to

$$Z=\sqrt{X^2+Y^2}\ \arctan\frac{Y}{X}$$

Echo is the second time a talker hears himself talk or a listener hears the talker due to delay in transmission time. The delay experienced on a physical facility may be as low as 14,000 mps/mi (mps per mile), whereas delay on carrier-derived facility approaches the speed of light. Talker echo is due to an impedance irregularity at a distant point, whereas listener echo is usually due to an impedance irregularity at or near both ends of a telephone connection. Echo range is considered to be between 500 and 2,500 cps (cycles per second).

Echo is a function of the reflection coefficient p, and

$$p=\frac{Z_2-Z_1}{Z_2+Z_1}\qquad Z_2>Z_1$$

where Z_1 is line impedance and Z_2 is network or reference impedance, or vice versa.

Echo or reflection coefficient is difficult to measure directly; therefore, return loss measurements are generally used. We can assume that echo limits are met when all switched tandem trunks meet return loss design objectives. Otherwise, echo may be limiting.

Return loss measurements indicate the degree of match or image impedance existing at the junction of two complex elements and is equal to

$$20\log_{10}\frac{Z_2+Z_1}{Z_2-Z_1}\qquad Z_2>Z_1$$

In addition, return loss is equal to two times the normal line attenuation as the signal traverses the same path out to an impedance irregularity or junction and then back to the testing end. To summarize, return loss values at a reference point are combined on a power summation basis, whereas the effect of an impedance dissimilarity at a distant point is removed by adding two times the net loss to the distant point; the sum is then reduced only by the structural return loss of the interconnecting facility added on a power summation basis.

Singing results when circuit net losses equal gains provided. Actually, marginal-to-singing condition causes hollowness and poor transmission characteristics when less than 6 db. A margin from singing of at least 10 db between 250 and 3,200 cps for any voice-frequency repeater or carrier-derived connection is considered adequate.

Adding gain devices to overcome facility net losses incurs echo, return loss, and singing problems. Echo, return loss, and singing objectives are met by achieving an adequate degree of impedance match throughout any switched connection. The degree of impedance mismatch permitted yet meeting minimum return loss design objectives is illustrated by Fig. 2(A) for a 900-ohm rated office and Fig. 2(B) for a 600-ohm rated office. Of course, this means meeting or exceeding

Table I. Approximate Characteristics of Nonquadded Exchange-Type Cable at 1,000 CPS

Gage	Loop R/Mi	Mutual Capacitance, μf/Mi	Impedance† R	Impedance† X	Impedance† α	Impedance† θ	Cutoff Frequency, Cps	Loss, Db/Mi, 68 F	Distortion Penalty, Db/Mi
				A—Nonloaded					
26	440	0.079	672	−J660	942	44.5		2.86	0.57
24	274	0.084	517	−J503	721	44.2		2.31	0.46
22	171	0.082	416	−J399	576	43.8		1.79	0.36
19	85	0.084	295	−J273	402	42.8		1.25	0.25
19	85	0.066	333	−J308	453	42.8		1.11	0.23
				B—H-88 Loaded					
26	440*	0.079	1,114	−J423	1192	20.8	3,600	1.8	0.0
24	272*	0.084	1,039	−J271	1074	14.6	3,500	1.23	0.008
22	171*	0.082	1,036	−J177	1031	9.7	3,500	0.79	0.025
19	85*	0.084	1,013	−J92	1017	5.2	3,500	0.42	0.07
19	85*	0.066	1,132	−J103	1137	5.2	3,900	0.374	0.0

* Add 5.4 ohms for each load point.

† Mid-section iterative Z in cases of loaded facilities.

F = Degrees Fahrenheit.

the return loss design objective over the singing range, namely, 250 to 3,200 cps (which includes the echo range).

Fig. 3(A) illustrates the current and possible future direction of return loss design parameters and (B), the method used in its development. Return loss as used in Fig. 3(A) and (B) is that "seen" when looking directly towards the next lower class office(s) connected through to the subscriber's set. The effective return loss is the calculated net resultant of a switched-through connection; see Fig. 3(B).

As may be seen by the solid curve in Fig. 3(A), the return loss design parameters of 12 db at class 5 end offices, 15 db at class 4 toll centers, and 27 db at all remaining higher class offices just barely meet the minimum effective return loss objective shown by the dotted line. Improving upon the class 5 office return loss objective results in only a nominal overall improvement, as illustrated by the dash-dot line in Fig. 3(A). Improving the office return loss objective at a class 4 office would provide a comfortable, effective return loss margin, as shown by the dashed line in Fig. 3(A). Such an improvement will permit reducing the toll-connecting trunk net loss design objectives which is, no doubt, the next step in improving upon transmission levels. Fortunately, no real difficulty is encountered in upgrading (to improve upon return loss figure) class 4 offices.

Basic Impedance Characteristics of Interconnected Facilities

Since return loss of local plant and connecting trunks play such an important part in attaining and maintaining good transmission, let us consider some of the impedances which determine the return loss value encountered. Fig. 3(A) points out the need for adequate return loss. Any appreciable reduction in return loss at any one of the switched offices will frequently result in adverse subscriber reaction.

Typical characteristic impedance of a nonloaded cable pair at 1,000 cps only is provided in Table I(A); Table I(B) shows an appreciable change in characteristic impedance when *H88*-type loading is applied to the same cable pairs, also shown for the 1,000-cps frequency point only. However, Table I does not provide sufficient information to permit adequate matching of impedances over the singing frequency range. This is readily seen by comparing Fig. 4, which illustrates characteristic impedance of nonloaded cable pairs, with Fig. 5 illustrating possible impedance characteristic of a loaded cable pair for various lengths of end section and across the voice frequency band.

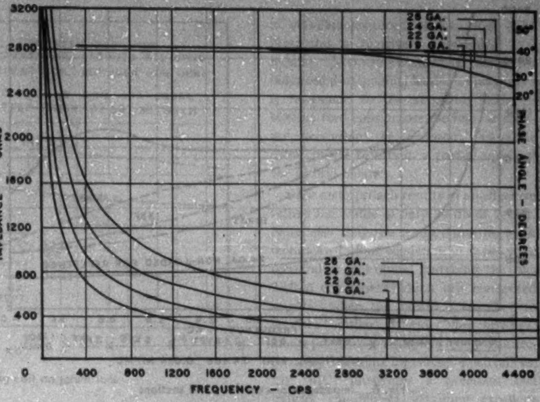


Fig. 4. Typical impedance characteristics of nonloaded cables

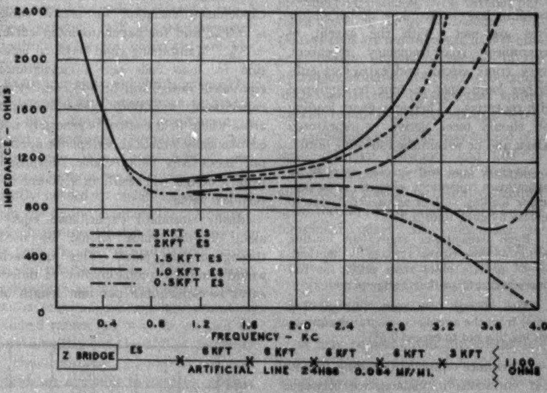


Fig. 5. Impedance of loaded cable for various end sections

Loading a cable pair is merely a means of effectively increasing the inductance of a nonloaded cable pair so as to counter or balance out its inherent capacitance per unit length. As a result, the nominal impedance of a loaded cable pair is higher than the equivalent nonloaded cable pair and presents mostly resistive reactance as its new impedance characteristic, reducing the net loss per unit length.

It is interesting to note in Table I(B) that the nominal impedance of 19-, 22-, and 24-gage (high-capacity) cable pairs is approximately the same and, since the *Q* of a resonant circuit is a function of series resistance, the cutoff characteristics of cable pairs of this gage remain the same. Also, the impedance characteristic below approximately 800 cps changes with the amount of series resistance encountered and is not due to resonance of a loading system; such resonant points are generally well above 3,000 cps. Because of this the loaded cable impedance characteristic below approximately 800 cps can be seen by moving up on the curves in Fig. 4 so that it passes through the correct impedance at 1,000 cps, taken from Table I(B), for the loaded cable involved.

Since loaded cable is being used quite extensively, design criteria for loading

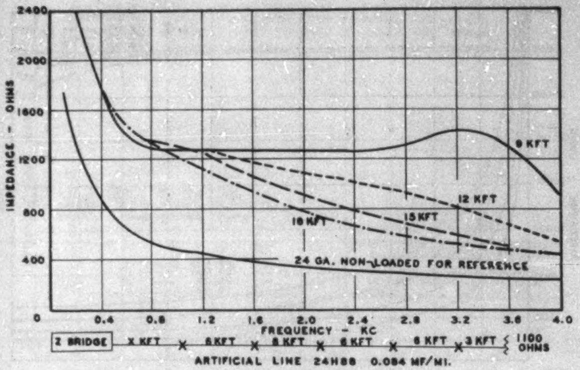


Fig. 6. Impedance of extra-long end sections

systems will be provided to help in understanding how return loss design objectives are met. With the advent of economical voice-frequency repeaters, every trunk circuit and extra-long subscriber loops may in time be equipped with repeaters. Therefore, every loading job should meet active or repeatered design and be within the following limits:

1. Average load coil spacing as close to capacitance equivalent spacing for a particular loading system as practical, not to deviate by more than ±2%.
2. Each capacitance equivalent loading section as close to the average as practical, not to deviate more than ±2% for toll routes and ±3% for exchange plant.
3. The average departure (disregarding signs) from the average capacitance equivalent loading not to exceed 0.5%.

Capacitance equivalent loading means that variations in capacitance between load points meets the limits set forth above irrespective of changes in capacity per unit length of different cables encountered throughout the length of a route.

In a cable-loading design, care must be taken to consider differences in interwinding capacitance of load coils resulting from different manufacturing techniques. Typical standard-wound load coils present a bridging capacitance of 250 to 300 $\mu\mu$f (micromicrofarads). A bifilar-wound coil may add a 2,500- to 3,200-$\mu\mu$f capacitance. The difference in capacitance between the two coils can add the equivalent of a 3% difference in physical distance between load points of *H*-type loading system and 6% on *B*-type loading systems.

The inductance value of load coils is held within ±0.5% of standard values. From tests made by a cable manufacturer, capacitance variation for plastic cable is ±3%, and for paper-insulated cable, ±7%. Maintaining rigid physical location for load coils per a capacitance equivalent design will help in countering capacitance variations per-unit-length areas where field control is generally not economically practical, except for special purposes. Any appreciable deviation from a design will result in a poorer return loss value.

Ideally, standard capacitance equivalent loading value should be used throughout the plant. This approach would eliminate problems due to differences in capacitance per unit length of different cables. Also, such a practice would help materially in matching cutoff impedances at switching points.

Since the length of the end section plays such as important part in determining the terminating impedance of a loaded cable pair, the first load point out of an office should be not more than 0.45 times the average capacitance equivalent between load coils. Much of the remaining capacitance required for building out to 0.5 end section is provided by the office switches and associated cabling. Fortunately, the end section can be increased in length electrically simply by using a bridging capacitance for any remaining deficiency.

When an end section is too long, there is little that can be done except to build out to a full end section, adding a full-weight load coil and then to build out to the desired or 0.5 end section. Alternatively, it may be possible to use a one-half weight load coil and appropriately shorter end section to do the equivalent. Image impedance of loaded cable exists only at the mid-section (or 0.5 end section) and at the mid-point of a load coil. As information, end sections greater than 1.0 present impedance characteristics, as shown in Fig. 6. Obviously, long end sections would present a serious impedance mismatch and this, especially at switching points, is to be avoided.

Telephone sets act as a termination on the end of a subscriber's loop. Fig. 7 illustrates a typical telephone set of current design. The range of its characteristic impedance results from different values of loop current. The nominal im-

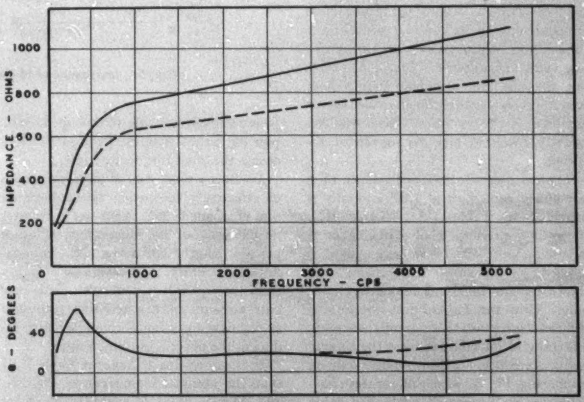


Fig. 7. Typical telephone set impedance characteristics

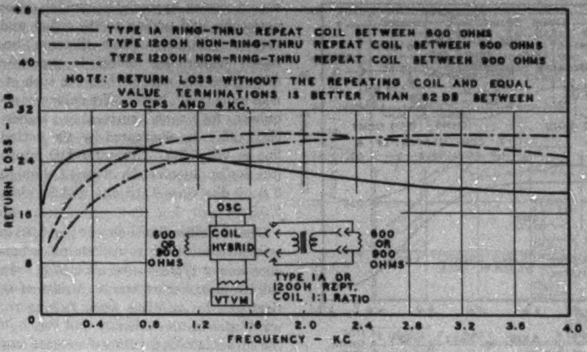


Fig. 8. Effect of single repeating coil on return loss

pedance is a compromise between nonloaded and loaded cable facilities achieving as a result not a particularly good impedance match of either. Possibly, future telephone sets will standardize on loaded cable facility impedance characteristic or will be provided with a means to match not only loaded but nonloaded cable pairs or 600-ohm compromise terminations.

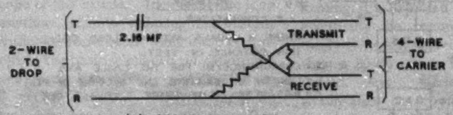
Repeating coils are used extensively for isolating circuits longitudinally and for matching impedances of different facilities. A trunk equipped with a repeating coil which can be switched to a coil hybrid with the use of a precision or compromise network on the net side normally requires the addition of a similar balancing repeating coil between the hybrid and its associated network, in order to attain adequate balance for class 3 and higher-class offices. Fig. 8 illustrates the degrading effect a repeating coil has on return loss when connected as shown and without an equal balancing unit on the network side of the hybrid. For this reason, repeating coils are avoided in the higher-class offices. Using repeating coils in classes 4 and 5 offices generally will not require extra attention as design objectives are less stringent.

Resistance and coil hybrids are used for converting 4-wire connections to 2-wire. Fig. 9(A) illustrates a resistance-type hybrid sometimes permitted at end offices only. Note that 2.16-μf condenser is in series with the line and is frequently needed to isolate direct current from the hybrid components. However, there is no equal balancing condenser in the network; this is mainly why return loss is relatively low. Under certain conditions when no direct current isolation is required, some improvement in switched return loss may be realized by short-circuiting out the 2.16-μf condenser in the resistance hybrid. Resistance hybrids are available in 600-ohm nominal impedance only, necessitating the use of 1 to 1.5 ratio repeating coils when installed in 900-ohm switched offices. Fig. 9(B) illustrates a relatively economical coil hybrid. No means is provided therein for balanced d-c signaling leads. This makes it necessary to use repeating coils or battery feed coils across the line for making accessible the necessary signaling leads, and it results in a partial upset of the hybrid balance.

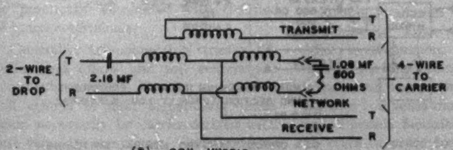
Such an approach results in an effective return loss value of only 20 db or better. Note also that the compromise network includes a 1.08-μf condenser in series with (generally) 600 ohms, matching the two 2.16-μf condensers which are connected in series when the hybrid is switched to another similar hybrid. This arrangement provides a relatively good balance, certainly adequate for classes 4 or 5 offices. Fig. 9(C) illustrates a precision coil hybrid, capable of attaining excellent balance required in class 3 and higher-class offices and incorporating means for providing d-c signaling paths without partially upsetting the hybrid balance. Repeating coils are not required, as 600-ohm and 900-ohm impedance versions of the hybrid are available. From the foregoing discussion, it may be seen that adding hybrids requires some attention to detail.

Compromise networks may consist of a series 600-ohm resistance and 1.08- or 2.16-μf capacitor, or 900-ohm resistance and 1.08- or 2.16-μf capacitor, depending upon the impedance standardized for the individual office. The discussion on hybrids pointed out various compromise network values encountered. The need is apparent for standardization within an office in order to attain desired return loss values.

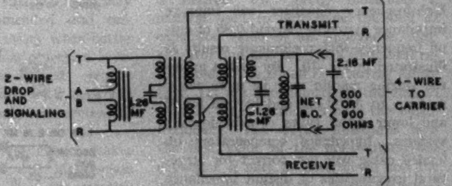
As may be seen from Fig. 5, the impedance characteristic of loaded cable approaching cutoff rises rapidly. Since compromise networks present essentially a flat impedance characteristic well be-



(A) RESISTANCE HYBRID



(B) COIL HYBRID



(C) PRECISION COIL HYBRID

Fig. 9. Typical hybrid configurations

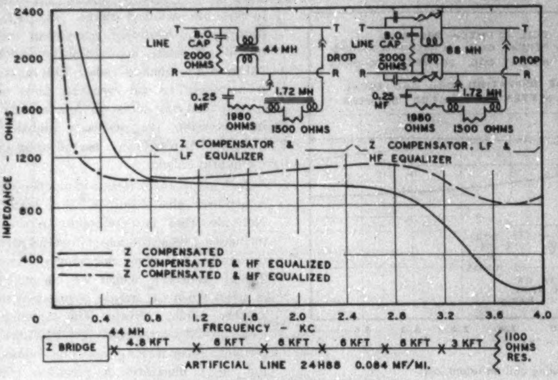


Fig. 10. Effect of L compensated and LF equalized

yond cutoff, either the compromise network needs to be modified to resemble a loaded facility (precision network) or the loaded cable impedance must be held flat up to at least 0.85 of its cutoff frequency simply by adding an impedance compensator. An impedance compensator consists of a one-half weight load coil with cable end section building-out capacitors. It is primarily effective above 1,000 cps. The building-out capacitor is used to increase electrically the cable end section to about 0.8 of a full section.

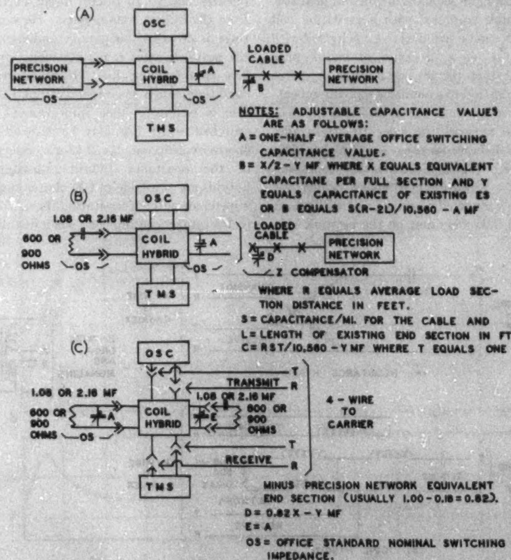
At the 0.8 end section, the compensated cable impedance characteristic is essentially flat and primarily resistive from 1,000 cps to 0.85 of cutoff, as shown on Fig. 10. Since an appreciable low-frequency impedance mismatch exists between 19- and 24-gage cable pairs (a modification of Fig. 4 as discussed earlier), a low-frequency compensator is added on 24-gage facilities to make it look electrically similar to 19 gage when necessary. Although an impedance compensator (see Fig. 10) will do a fair job, the addition of a high-frequency equalizer will improve impedance matching beyond 0.85 of cutoff frequency with preliminary results as illustrated in Fig. 10. By such a means the singing margin may be more easily met up to 3,200 cps and beyond for *H* and lower cutoff types of cable-loading systems. Impedance compensators on loaded cable will not correct for structural irregularities or distant impedance mismatches.

Matching Impedances

As noted earlier, switching together image impedances is desired for attaining minimum echo and reflection losses. Fig. 3 illustrates the need and the magnitude required for adequately matching impedances at switching points.

The minimum effective return loss objective on any switched subscriber-to-subscriber connection is 10 db. For short distances, as in local switched connections, echo, singing, noise, and crosstalk are not particularly significant, as transit time and exposure are limited. However, as distances increase, transit time becomes appreciable, making echos noticeable and adding to the talker and listener problems. Since distance between higher-class offices is generally quite great and net loss between such offices quite small, better office balance is required for each succeeding higher-class office, as illustrated by the dotted line in Fig. 3(B). Current DDD return loss design objective for classes 1 through 3 is 27 db; class 4, 15 db; and for class 5, 12 db.

The end office return loss in Fig. 3(B) is estimated at 12 db to include office balance and a typical subscriber loop. In this example, return loss is considered as that seen by an office when looking towards the distant office(s) and "seeing" the subscriber's set. Effective return loss is the actual return loss seen when switched through an office and looking towards the distant office(s) to the subscriber's set. When minimum return loss for each class of office exists, the class 4 office just meets minimum effective return loss objective. Improving upon office balance at class 4 offices is indicated. Such an improvement will permit reducing somewhat the net loss between classes 5 and 4 offices, the source of the major remaining loss in a toll connection. Any appreciable improvement thereafter is then limited by the end office.



NOTES: ADJUSTABLE CAPACITANCE VALUES ARE AS FOLLOWS:
A = ONE-HALF AVERAGE OFFICE SWITCHING CAPACITANCE VALUE.
B = X/2 − Y MF WHERE X EQUALS EQUIVALENT CAPACITANE PER FULL SECTION AND Y EQUALS CAPACITANCE OF EXISTING ES OR B EQUALS S(R−2L)/10,560 − A MF WHERE R EQUALS AVERAGE LOAD SECTION DISTANCE IN FEET.
S = CAPACITANCE/MI. FOR THE CABLE AND
L = LENGTH OF EXISTING END SECTION IN FT.
C = RST/10,560 − Y MF WHERE T EQUALS ONE MINUS PRECISION NETWORK EQUIVALENT END SECTION (USUALLY 1.00 − 0.18 = 0.82).
D = 0.82X − Y MF
E = A
OS = OFFICE STANDARD NOMINAL SWITCHING IMPEDANCE.

Fig. 11. Return loss and lineup testing

Subscriber loops are generally based on theoretical natural loop resistance design. This design calls for loading all subscriber loops longer than 20 kft (kilofeet) in distance. A facility approaches its characteristic impedance when its length is sufficient to incur more than 3 db of line loss. A facility long enough to incur over 6 db of line loss presents its characteristic impedance. Since most common cable facilities used on subscriber loops incur more than 3-db line loss before approaching the 20-kft limit, it follows that we are switching loaded and nonloaded subscriber lines to toll-connecting trunks which are generally loaded cable or carrier-derived facilities. This explains one important reason for return losses being so low at end offices.

We have seen how loaded cable end sections can be so varied as to match an image impedance. Switching together two half end sections will then resemble continuous loading. Improving end office return loss on nearly all subscriber loops can be accomplished simply by loading all cable facilities the length of which exceeds 3 db in loss at 1,000 cps even though only one or two load points may be required. Providing telephone sets with impedance characteristic more nearly equal to loaded cable will materially improve end office return loss and help to improve all transmissions.

Incidentally, it must be apparent that further improvements in existing telephone set efficiencies will not be needed as loading provides the added efficiency required to meet current and immediate future subscriber transmission level objectives. This is good, as the current telephone set efficiency has reached its peak and any improvement in the transmitter or receiver individually will result in only a very nominal effective improvement. The procedure is more complex than simply $T+R/2$ and is beyond the scope of this article. Increasing the loop resistance, thereby reducing line current available, would mean using another concept in the basic telephone set denoting a breakthrough most likely requiring the use of active elements. Bridged taps, an important reason for many subscriber transmission complaints, will eventually be practically eliminated with the use of junction terminals.

Generally, toll-connecting trunks between end offices and toll centers are physically derived. Trunks to adjacent toll centers or next higher-class offices must necessarily be carrier-derived with but very few exceptions. This is necessary as carrier-derived return losses are limited only by terminating coil hybrids and their networks. As a rule, physical facility return losses are only 19 db or better, certainly not adequate when an attempt is made to meet 27-db return loss objective.

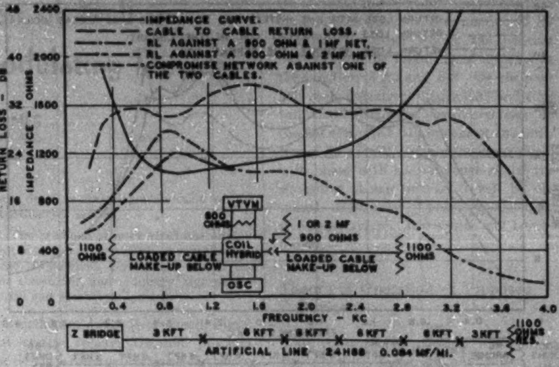


Fig. 12. Return loss of loaded cable, one half end sections

Standard Office-Switching Impedances

Several approaches exist as to what switching impedance shall become standard for any single office. Local exchange switching for economic reasons invariably will be so arranged as to switch on half end sections of loaded subscriber lines. Nominal impedance of loaded cable facilities is a little over 900 ohms. For lines not loaded and yet physically short enough to incur less than about 3 db of loss, the impedance seen will be primarily that of the telephone set. For lines physically longer and incurring more than a 3-db net loss which have not reached approximately 20 kft in length, the point at which theoretical natural loop design calls for loading, it is expected that load coils may eventually be recommended and certainly they will be necessary when better than a 20-db return loss objective is required. Any carrier-derived trunks would then be provided with networks simulating normal half end section of loaded cable.

Future expansion via carrier makes switching at class 4 and higher-class offices most common. Since the quantity of trunks is large, nominal office switching is based on compromise networks. Impedance compensators must then be used on all loaded facilities coming into such an office in order to meet and improve upon current return loss design objectives.

Test for Image Impedances

To many in the telecommunications field, the foregoing would appear to be a difficult and complex problem; however, a simple means for determining the suitability of any facility is readily available. In the past, most impedance-matching problems involved tedious and time-consuming impedance measurements and calculations in order to attain the desired impedance characteristics. Although such may still be required for locating some troubles on a physical facility, the adjustment of end sections, office balancing, and determining suitability of a facility can readily be determined simply by making return loss measurements against known standards.

Fig. 11(A) illustrates a method making acceptance and structural return loss measurements on a loaded-cable facility. Such a test can be used to adjust the end section building-out capacitor *B* when approximately half end section switching is desired. Fig. 11(B) illustrates a method used to determine the best *D* capacitor value of an impedance compensator and also measures structural return loss. The *B* and *D* capacitor values (Fig. 11) are determined by the test to provide best return loss near the cutoff frequency. Fig. 11(C) illustrates how the best value of the net building-out capacitor *E* may be determined and the compromise network values checked against an office "standard" termination.

Return loss of a switched-through loaded-to-loaded facility based on half end sections is limited only by structural return losses, as illustrated by Fig. 12, which also shows the results when a compromise network is matched against a

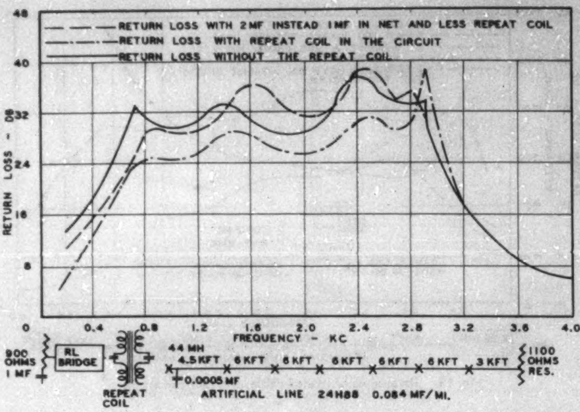


Fig. 13. Effect of repeating coil and impedance compensator on return loss against compromise network

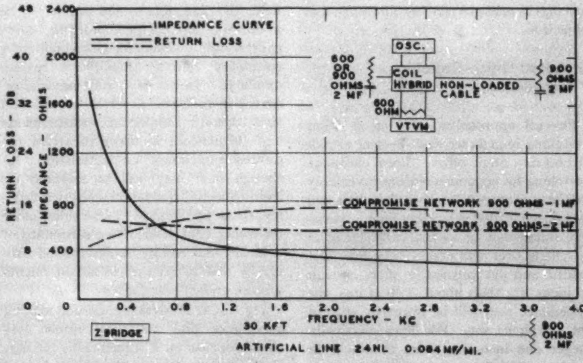


Fig. 14. Return loss of nonloaded cable against compromise network

loaded line. Fig. 13 illustrates the improvement in impedance matching attained when a compromise network looks into a compensated loaded cable. Adding a low- and high-frequency equalizer would have improved the low- and high-frequency return losses.

Conclusions

Controlling echo and singing margin will assure good voice-frequency and data transmissions. Echo and singing are controlled by an adequate degree of match between impedances of switched facilities to include the connecting trunks. The degree of match required must be adequate to permit adding sufficient gain in a circuit to overcome as much of the line losses as practical.

Return loss measurements provide the resultant when two complex impedances are combined. Meeting return loss design objectives will assure adequate echo and singing margins.

Via fundamental planning, all facilities within an office can be made to match a "standard" for an office. An office standard may be a compromise network with bridging capacitance equal to one-half the average office-switching capacitance. Such a standard can be used in offices having compensators on loaded cable pairs and compromise networks on coil hybrids. Alternatively, a precision network built out to resemble a half end section can become a standard in offices switching on half end sections. For best results such an office should have all cable pairs, equivalent capacitance loaded to a single value. Based on return loss measurements and using the office standard, all loaded cable facilities would be built out capacitively to a value indicating the best return loss needed above 1,000 cps. Whether low-frequency equalizers are needed will depend upon whether 24-gage cable is to be matched with 19 gage within the office under consideration.

The effect that various parts of a telecommunications network have on return loss is demonstrated by Figs. 8, 10, and 12 through 14. This is aimed at pointing out fundamentals involved in matching impedances and technique for making tests. Generally, adequate impedance matching requires engineering effort, accurate records, and only a nominal amount of extra equipment. The end result can be most gratifying.

The tedious job of making point-by-point measurements and graphs is all but eliminated when a swept frequency oscilloscope transmission measuring set is used in lieu of the oscillator and transmission measuring set shown in the figures. Such visual plotting will immediately display the result of any changes in a circuit. This not only saves appreciable time but, more important, will permit optimizing terminating impedances and voice-frequency repeater gains or measure structural return losses (acceptance testing) as a matter of routine by relatively inexperienced personnel and with a minimum of instruction.

When facility net loss measurements are made with a swept frequency transmission measuring set, the sending end will display an approximate impedance curve while the receiving end sees the frequency response curve. Repeater gain, approximate impedance, return loss, net loss measurements, etc., can be readily made with the new tester.

References

1. Study of Return Losses on Loaded Trunk Cables and Methods Taken to Improve Them, **G. H. Speake.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, May 1958, pp. 201–04.

2. Nation-wide Switching of Intertoll Trunks—Transmission Considerations, **O. F. Wallman.** *Ibid.*, pp. 223–30.

3. Communication Networks and Lines (book), **Walter J. Creamer.** Harper and Brothers, New York, N. Y., 1951.

4. Electrical Communications Experiments (book), **Henry R. Reed, T. C. G. Wagner, George F. Corcoran.** John Wiley & Sons, Inc., New York, N. Y., 1952.

5. Circuit Analysis of Transmission Lines (book), **John L. Stewart.** John Wiley & Sons, Inc., 1958.

On the Elimination of Null in Modulating and Demodulating Devices

SHELDON JONES
ASSOCIATE MEMBER AIEE

Synopsis: A method of design is proposed for electromagnetic a-c to d-c and d-c to a-c conversion devices which largely eliminates the interdependence of bias voltage and dynamic range. This is achieved by using a model approach which divorces the output null from the dynamic range. Test results have proved the method achieves the desired results without affecting the characteristics of the basic device.

THE USE of devices to convert alternating current to direct current and vice versa is of vital importance to the electronic designer, particularly in the field of servomechanisms. One problem that usually confronts the designer in modulating and demodulating devices is the one of null reduction. (Null is defined, for this paper, as the output voltage of a device referred to the input, for zero signal input.)

The null of modulating and demodulating devices is usually a function of the maximum signal to be handled by the device and is usually in the order of one per cent of the maximum over the temperature range −55 C to +85 C (degrees centigrade). The exception to this rule is mechanical choppers; however, the use of choppers is restricted by amplitude, power, temperature, reliability, and carrier frequency considerations.

One of the most common methods of null reduction is the use of a bias adjustment; see Fig. 1. However, null cancellation by the use of adjustments has the following disadvantages:

1. Adjustment can be made for only one ambient temperature, and therefore the null cannot be maintained over a range of temperatures.
2. Adjustments are subject to improper setting and inadvertent movement of adjusting potentiometer.
3. The output impedance of the associated phase detector may be a function of the potentiometer setting.
4. The presence of more than one adjustment in a system may result in complicated, cyclic adjustment procedures which are time consuming and require a high degree of skill and special equipment.

A method of design is proposed, Fig. 2, that greatly reduces the null problem. The design is discussed in terms of a phase-sensitive detector. But, by making the obvious changes the method is equally applicable to modulating devices. The technique could be used in phase detectors or modulators constructed with electronic, magnetic, or mechanical components.

Phase-Sensitive Detector Operation

The operation of the phase-sensitive detector portion of Fig. 3 will be explained. The circuit consists of *T1*, *T2*, *R3*, *R4*, *CR1*, *CR2*, *CR3*, and *CR4*. Transformer *T1* is necessary for converting single-ended signals to a balanced signal for full-wave operation. Resistors *R3* and *R4* are necessary for limiting current in the transformer, *T2*. Transformer *T2* provides the voltage for operating the switching diodes *CR1*, *CR2*, *CR3*, and *CR4*. The polarity relations of the windings are denoted by dots adjacent to the windings. A reference signal of the same frequency and phase locked with the input signal is used for the switching voltage.

Assume that a sinusoidal signal in phase with the switching voltage is impressed on the isolation transformer, *T1*. The secondary of transformer *T1* will have an out-of-phase signal on the dot end of the winding and an in-phase signal on the opposite end. This signal will appear on *T2* as shown by the circuit. The dot end of the windings on *T2* will then be in phase with the switching voltage.

Consider the instant of time when the reference signal is positive with respect to ground. Diodes *CR3* and *CR4* will conduct; *CR1* and *CR2* will not. Therefore, the in-phase end of transformer *T1* will be essentially coupled to the junction of *R3* and *R4*. In other words, the positive half-cycle of the input signal appears at the junction of *R3* and *R4*. On the next half-cycle (negative) diodes *CR1* and *CR2* will conduct and the positive voltage on the dot end of *T1* will appear at the junction of *R3* and *R4*. When the input signal is in phase with the reference signal the waveform shown in Fig. 3 will be obtained. For an out-of-phase signal a negative full-wave rectified signal will be obtained.

Theory of Operation of Proposed Device

By conventional methods, low null, in the absolute sense, is not compatible with large magnitude of signal capability. The method proposed makes use of the absolute null obtained by low-level signal device, but still retains large-amplitude capability, i.e., the null of the output member is essentially independent of dynamic range.

The proposed method of achieving this is shown in Fig. 2. The output of the phase-sensitive detector is compared with the input signal. This action will correct the error in the output. For no input, the bias of the phase detector will be reduced by the loop gain of the feedback loop.

One way of illustrating the null reduction is to reduce Fig. 2 to Fig. 4 by using standard block diagram techniques. From Fig. 4 by inspection:

$$E_0 = E_{in}\frac{(1+K_2K_3K_4)K_1}{1+K_1K_2K_4} + \frac{\Delta V_1}{1+K_1K_2K_4} - \frac{\Delta V_2K_4K_1}{1+K_1K_2K_4} \qquad (1)$$

By making $K_1 = K_3$ (fixed gains in the circuit design) and $K_1K_2K_4 >> 1$, this reduces to

$$E_0 = E_{in}K_1 + \frac{\Delta V_1}{K_1K_2K_4} - \frac{\Delta V_2}{K_2} \qquad (2)$$

This equation can be obtained just as easily from loop equations. Equation 2 demonstrates that the bias ΔV_1, bias due to the phase detector, is divided by the loop gain $K_1K_2K_4$ and a new bias term, ΔV_2, has been introduced. The voltage required to give zero output from the modulator is $\Delta V_2/K_2$. This voltage can

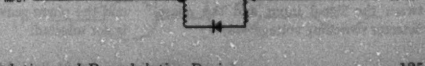


Fig. 1. Phase detector with adjustment

Paper 61-25, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 23, 1960; made available for printing November 16, 1960.

SHELDON JONES is with Hughes Aircraft Company, Culver City, Calif.

The author wishes to thank James B. Hayes, formerly of Hughes Aircraft Company, for the building and testing of the circuit in this paper.

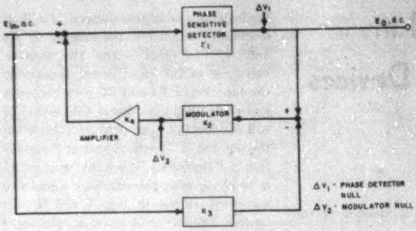

E_{in}, a.c.
+
-
PHASE SENSITIVE DETECTOR K_1
ΔV_1
E_0, d.c.
+
-
MODULATOR K_2
K_4
AMPLIFIER
ΔV_2
K_3
ΔV_1 - PHASE DETECTOR NULL
ΔV_2 - MODULATOR NULL

Fig. 2. Block diagram of low-null phase detector

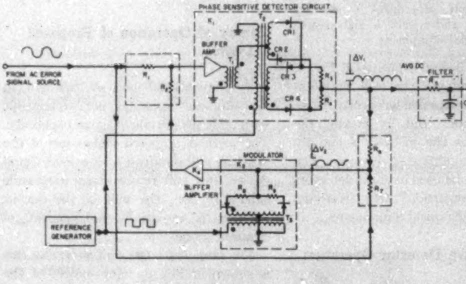


Fig. 3. Circuit diagram for phase detector

be kept small since the output of amplifier K_4 is only required to be greater than $\Delta V_1/K_1$. Therefore, the range of the signal the modulator is required to handle is $\Delta V_1/K_1K_2K_4$. Theoretically this can be made as small as desired.

The summation on the input to the modulator is an a-c to d-c summation. Fig. 5 shows the different waveshapes. Fig. 5(b) is the unfiltered waveshape on the output of the phase detector. Fig. 5(a) is the sinusoidal input waveshape. Fig. 5(e) shows the waveform across the modulator. From zero to π the summing point is grounded. From π to 2π the modulator represents a high impedance to ground and the summing occurs giving the waveshape of Fig. 5(e). In practice, it would probably be necessary to add some filtering for loop stability of the phase-detector output. Therefore, the waveshapes of Fig. 5(b) and (e) will be somewhat different.

A factor of fundamental importance is whether the phase-sensitive detector will exhibit the characteristic of phase resolution. Consider a sine-wave input to a phase detector of the form:

$$E_{in} = E \sin(\omega t + \varphi) \qquad (3)$$

where φ is an arbitrary phase angle between the signal input and the phase-detector switching voltage.

Equation 3 can be restated as

$$E \sin(\omega t + \varphi) = E(\sin \omega t \cos \varphi + \cos \omega t \sin \varphi) \qquad (4)$$

Equation 4 shows the well-known fact that a sine wave with displacement is the sum of a sine and cosine wave whose amplitudes vary as a function of the phase displacement.

The detector will reject the $\cos t$ term because the average value of the cosine term is zero over the interval 0 to π. Thus, it is seen that the unfiltered output, E_0, will be $E \sin \omega t \cos \varphi$ where $\cos \varphi$ is a constant which affects only the amplitude of the output. The same signal, equation 4, will be on the input to the modulator through K_3.

Since the modulator is open and short-circuited in alternate intervals of π, see Fig. 5, no signal will exist at the modulator when the summing point is at ground potential. During the summing interval π to 2π, the average value of the cosine term is zero and the only contribution at summing is the $\sin \omega t \cos \varphi$ term.

Therefore, it is obvious that both terms, summed, are sine terms whose amplitude varies as a function of $\cos \varphi$. Since both terms vary as $\cos \varphi$, both exhibit the same phase discrimination properties, and the phase sensitivity of the detector is not affected.

The negative sign used at the summation point, Fig. 2, for the a-c input to the modulator arises from the phasing chosen for the modulator, which determines the half-cycle of the a-c input to be summed.

Description of Circuitry

The circuit* described herein is a phase-sensitive detector-stabilized circuit with a feedback loop as shown in Fig. 3. The function of the feedback loop is to eliminate the adjustment required to obtain a zero null, and to maintain this null over a temperature range of −55 C to +85 C. The device was constructed with transistors and vacuum tubes as the amplifying devices, and diodes for modulating and demodulating. The problems involved in the design of the phase detector are directly related to the requirements. In this case, the requirements were as follows:

1. Temperature range of operation −55 C to +85 C.
2. ±20-volt phase detected output.
3. 5-millivolt null over temperature and frequency range.
4. Frequency of operation range of 10 to 1.

It has been found that a loop gain of 200 would be adequate to maintain a null of 5 millivolts for a normally functioning phase detector. This required a gain of approximately 400 in the feedback path. However, this brought about a fundamental problem. The null output is normally a relatively small amount when compared with the large signal output of the phase detector. Thus, with the exception of very-low-level error signals, the feedback circuit should have little effect on the phase-detected output. This may be accomplished with little difficulty for low-level signals. However, as the signal increases, the amplifier may begin to clip and distort the feedback signal. The level at which this clipping begins is a function of gain requirements and operating limitations imposed on the amplifying device. If the amplifier is permitted to clip asymmetrically, the phase-detected signal is altered.

Another problem concerned the phase detector alone. To obtain a phase-detector output of 20 volts, it was necessary to step up through a push-pull phase-detector input transformer. This transformation was necessary due to the voltage limitations on the transistors used in the circuit. However, this transformation also placed a restriction on the load seen by the phase detector. The com-

---

* The circuit was built and tested by James E. Hayes, formerly of Hughes Aircraft Company.

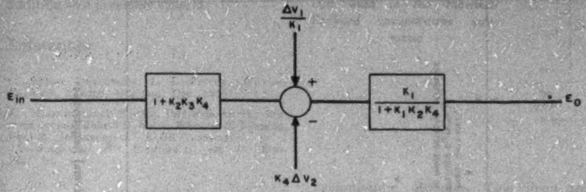


Fig. 4. Reduced block diagram of Fig. 2

bined load impedance of the feedback circuit and the output circuit could not be less than 100 kilohms.

The phase detector was a full-wave phase-sensitive detector with a gain of 0.9 d-c out to rms input. The primary of transformer *T2* was driven by a 20-volt peak-to-peak square wave. The turns ratio from the primary to the phase-detector secondaries was 5:1 stepup, and to the modulator secondary 10:1 stepdown. The phase-detector secondaries were tapped to 0.5% and the modulator to 0.1%. The maximum phase shift specified between the primary and phase-detector secondaries was 2 degrees with 1-degree maximum phase shift specified between the two windings. The maximum phase shift between the primary and secondary winding of the modulator, *T3*, was 2 degrees. Silicon quick-recovery diodes were used in the modulator.

The compensation network in the feedback path was designed to give a high d-c gain with a closed-loop frequency response of approximately 30 cps (cycles per second). The response was necessarily limited to this frequency by the inherent frequency limitation of the modulator carrier frequency.

The compensated output of the phase detector was summed with the input signal. The resultant was then modulated and amplified. However, an emitter follower has been inserted between the amplifier and the modulator to provide a high impedance sum point to keep the loss in the loop small. Special consideration has been given the modulator to keep the zero signal below 2 millivolts rms at the modulator. The diodes were matched for a current range of 0.9 ma (milliampere) to 1.1 ma, and the current-limiting resistors were 0.1%.

The curves showing the phase-detected output at high levels indicate that symmetrical output was obtained.

Test Results

The input-output characteristics of the device are seen in Fig. 6. The gain of the phase detector with feedback changed at high levels because the summing ratio on the modulator was not identical to the open-loop phase-detector gain.

The phase-discrimination characteristics of the detector with feedback at two different frequencies compared with an ideal cosine wave and with the phase detector alone are found in Fig. 7. The phase-discrimination characteristics of the two frequencies closely approach an ideal cosine wave shown in the same figure. The null of the phase detector without feedback has been taken into account in plotting the curve. On one device with feedback it was found that 12- to 15-degree phase shift on the signal was required to approach the ideal cosine wave, i.e., to get zero output at 90-degree phase shift, in the frequency range from 200 to 2 kc per second.

The results of tests on this device over the temperature range of −55 C to +85 C have been consistent with predicted performance. Nulls of less than 4 millivolts have been obtained over the temperature range for phase detectors having an output amplitude range of ±20 volts d-c. Stated differently, a dynamic range of 5,000 to 1 has been obtained over the temperature range.

Conclusions

The use of the proposed method results in improved performance in the following areas:

1. No adjustments are required.
2. Null reduction is achieved over a large range of temperatures.
3. Null reduction is achieved over a large range of carrier frequencies.
4. Changes in loading can be minimized.
5. Gain of the devices may be more closely held from unit to unit.
6. The null is essentially independent of the dynamic range of the device.

The device is not proposed as the solution for all null problems. However, if design requirements justify its use, the

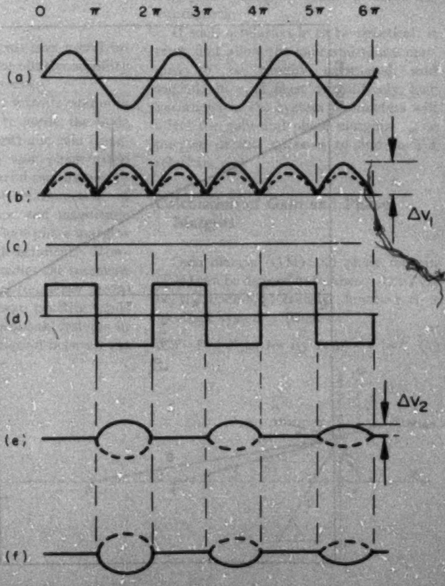


Fig. 5. Waveshapes of Fig. 2

a—A-c error signal
b—D-c output signal
c—Input voltage signal
d—Switching voltage for T2 and T3
e—Feedback signal out of modulator
f—Negative feedback at output buffer amplier

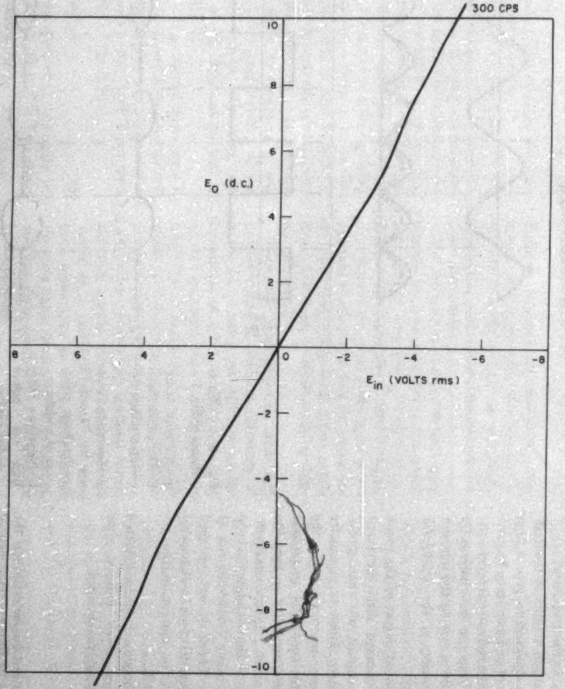


Fig. 6. Transfer function of a breadboard model

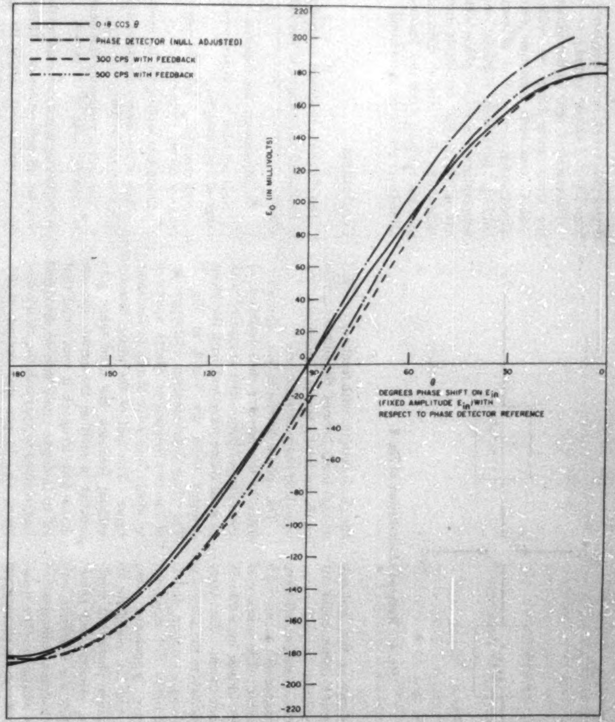


Fig. 7. Phase-discrimination characteristics for several frequencies

proposed method could be a valuable addition to a precision system.

References

1. Phase Detectors, L. I. Farren. *Wireless Engineer*, London, England, Dec. 1946.

2. Magnetically Keyed, Phase-Sensitive Demodulators, Robert B. Mark, Wheeler X. Johnson, Paul R. Johannessen. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 76, Mar. 1957, pp. 1–6.

◆

Discussion

Gerald Weiss (Polytechnic Institute of Brooklyn, Brooklyn, N. Y.): The results obtained by the author are interesting. It is rather unexpected that an improvement in null is obtained despite the addition of the modulator in the feedback return path.

I am wondering whether the author could compare his circuit with the well-known circuit used in the Hewlett-Packard vacuum-tube voltmeter. This circuit does not require a feedback modulator; a version is shown in Fig. 8.

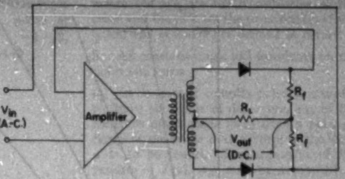


Fig. 8. Block diagram of circuit used in the Hewlett-Packard vacuum-tube voltmeter

Sheldon Jones: I would like to express my appreciation for Professor Weiss's interest in this paper. Since the null of the over-all device is determined by the modulator, the dynamic range of the modulator is kept very small. For large-signal operation the modulator is saturated and does not affect over-all device operation. By using the concept of obtaining the basic null of a small-signal modulator while retaining large-signal capability, the null of the over-all device is improved.

The author is not familiar with the Hewlett-Packard vacuum-tube voltmeter. The version shown in Fig. 8 of the discussion is a full-wave rectifier but is not a phase-sensitive detector. It would appear that the output voltage of this device would not be a function of a reference voltage in either phase or frequency as is the device discussed in the paper.

An attempt to feed back the voltage across *R3* and *R4* of Fig. 3 of the paper, as is done in the voltmeter circuit, would insert the full switching voltage on the input. The minimum value of this voltage would normally be at least twice the value of the input signal which would saturate the device.

A Method of Calculating Gain and Phase Margins From Pole-Zero Diagrams

F. H. SCHLERETH
NONMEMBER AIEE

Gain margin and phase margin are important quantities in the description of the performance of feedback systems since they are an accurate measure of relative stability. Their usefulness lies in the fact that they are easily calculated and that their variation with changes in system parameters is easily determined.

However, it must be noted that gain and phase margins are only applicable in the description of unconditionally stable systems whose open-loop transfer functions have no zeros in the right-half plane and whose open-loop characteristics are controlled by, at most, a single pair of complex poles. Since a majority of systems fall into this category, at least to a good approximation, the concepts of gain and phase margins give useful information concerning the relative stability of many practical systems.

In the synthesis of control systems in terms of the frequency response, the Bode Diagram [plots of logarithmic gain (magnitude) versus log ω and phase angle versus log ω], is a powerful real-frequency design tool. With the construction of one diagram, the phase and magnitude response, and the gain and phase margins of a system can be determined. However, it is difficult to predict the transient response with these real-frequency methods. Thus current design practice tends toward synthesis of feedback systems in the *s*-plane since the relation between the poles and zeros of a transfer function and the system transient and frequency response is well known.[1] It is, however, still useful to retain the notion of relative stability, and thus a relation between the positions of the poles and zeros of the open-loop transfer function and the gain and phase margins would be most advantageous.

If such a relation is to be practical, it must first allow the gain and phase margins to be rapidly calculated, and second, show, at least qualitatively, how variations in the system parameters will affect the gain and phase margins. The purpose of this paper is to find such a relation.

Calculation of Gain and Phase Margins

Gain margin (GM) and phase margin (PM) can be defined as follows. Let $T(s)$ be the open-loop transfer function of a feedback system. Then

$$GM = 1 - |T(j\omega)| \text{ for } \arg T(j\omega) = -180° \quad (1)$$

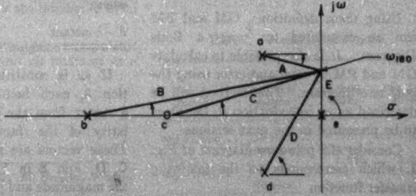


Fig. 1. Pole-zero diagram

Paper 60-1290, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960. Manuscript submitted June 17, 1960; made available for printing October 11, 1960.

F. H. Schlereth is with the General Electric Company, Syracuse, N. Y.

The author wishes to thank Dr. S. Seshu of Syracuse University for his valuable suggestions.

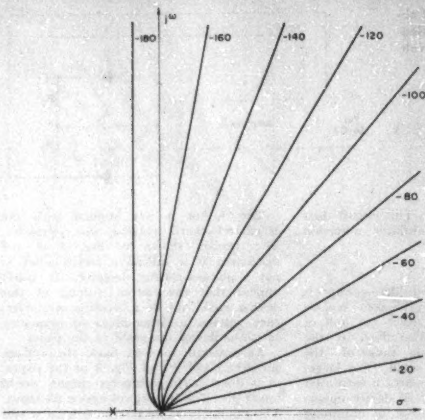


Fig. 2. Phase-angle loci for two poles on the real axis

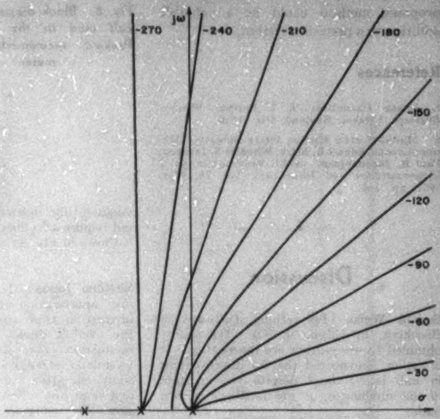


Fig. 3. Phase-angle loci for three poles on the real axis

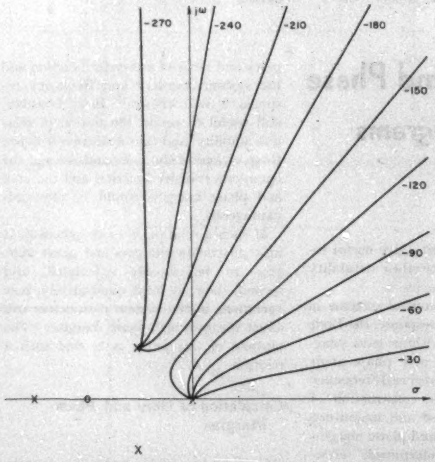


Fig. 4. Phase-angle loci for two complex poles, a pole at the origin, and a pole and zero on the real axis

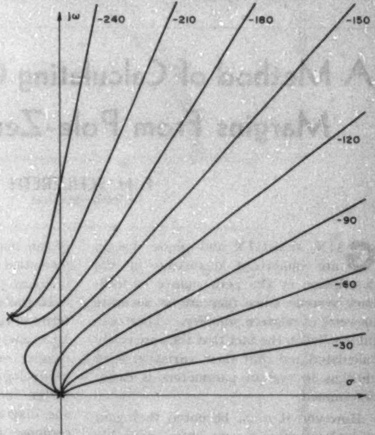


Fig. 5. Phase-angle loci for two complex poles and a pole at the origin

and

$$PM = \arg T(j\omega) - 180° \text{ for } |T(j\omega)| = 1 \quad (2)$$

Using these definitions, GM and PM can be calculated by using a Bode Diagram. It is also possible to calculate GM and PM by trial and error using the pole-zero diagram. This method will be described as an introduction to the ideas to be presented in the next sections.

Consider the pole-zero diagram of Fig. 1, which corresponds to the following transfer function

$$T(s) = \frac{K(s+c)}{s(s+a)(s+b)(s+d)} \quad (3)$$

where

K = constant

$s = \sigma + j\omega$ = complex frequency

If $j\omega$ is substituted for s in equation 3, each factor will correspond to a vector from the position of the singularity of the factor to the $j\omega$ axis. These vectors are represented by **A**, **B**, **C**, **D**, and **E** in Fig. 1. Furthermore, the magnitude and phase angle associated with each factor is represented by the length of the vector, and the angle the vector makes with the horizontal, respectively. (Counterclockwise rotation corresponds to a positive angle.)

The GM can be calculated by picking a point on the $j\omega$ axis, ω_{180}, at which the phase contribution of all vectors totals −180 degrees. The GM is then:

$$GM = 1 - \frac{\mathbf{KC}}{\mathbf{BDEA}} \quad (4)$$

To find the phase margin, the frequency,

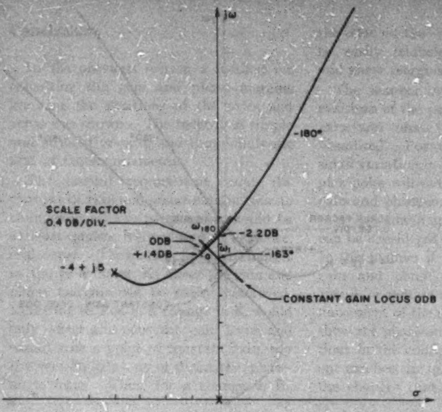


Fig. 6. Phase-angle loci for poles at $-4+j5$, $-4-j5$, $0+j0$

ω_1, corresponding to $T(s)=1$ must be found and the total phase contribution of the vectors calculated.

However, this method has the following disadvantages:

1. A trial and error procedure must be used to find ω_{180} and ω_1.
2. It is extremely difficult to see how variations in gain and shifts in the poles and zeros will affect the gain and phase margins.

Thus, although this is a relation between pole-zero diagrams and gain and phase margins, it is extremely impractical in terms of the discussion in the introduction.

Constant Magnitude and Constant Phase-Angle Loci

In the procedure for calculating the gain and phase margins by means of the pole-zero diagram presented in the previous section, the problem was reduced to finding ω_{180} and ω_1 by trial and error. This approach would be fruitful if a systematic procedure for determining ω_{180} and ω_1 could be found. With this in mind, the concept of constant magnitude and constant phase-angle loci will be introduced.[3] Constant phase-angle loci can be defined as images (in the s-plane) of radial lines in the $T(s)$-plane under the inverse mapping for $T(s)$. Similarly, constant magnitude loci are images of circles in the $T(s)$-plane under the inverse mapping for $T(s)$. Since the mapping is conformal, and since radial lines and circles form an orthogonal set of curves in the $T(s)$-plane, the constant magnitude and constant phase-angle loci will form an orthogonal set of curves in the s-plane.

Now, note that once these loci have been found, the points ω_{180} and ω_1 will be immediately determinable as the intersection between the -180-degree phase-angle locus and the $j\omega$ axis, and the zero db (decibel) magnitude locus and the $j\omega$ axis, respectively. Thus, these loci will provide a systematic procedure for finding ω_1- and ω_{180}. Now the practicality of determining the gain and phase margins from the pole-zero diagram hinges on the ease with which the constant magnitude and constant phase-angle loci can be calculated. Thus, at this point it is necessary to digress into a description of the construction of these loci.

Construction of Constant Magnitude and Constant Phase-Angle Loci

The construction of the constant phase-angle loci is greatly facilitated by the use of a spirule,* a device that enables arg $T(s)$ to be found accurately for any point in the s-plane. The following mathematical relationships are also extremely useful in the construction of the constant angle loci.[3]

1. The asymptote which is approached by the locus, as $s \to \infty$, for an angle ϕ is the straight line making an angle $-\phi/m-n$ with the real axis, where n is the number of finite zeros and m is the number of finite poles of $T(s)$.
2. If the asymptotes are extrapolated towards $s=0$, they will all intersect at a point given by

$$s_1 = \frac{\sum_{j=1}^{r} z_j - \sum_{k=1}^{s} p_k}{n-m}$$

where the z_j's are the co-ordinates of the zeros and the p_k's are the co-ordinates of the poles, and m and n are defined as before.

3. All constant angle loci terminate on the poles and zeros.
4. In the case of complex poles, it is useful to know the angle of departure, ϕ_D, of the constant phase-angle locus from a pole, p_1. Suppose the angle of departure of the locus ϕ is to be found. The argument of the residue of $T(s)$, ϕ_R, may be computed at p_1 and then the angle of departure is given by:

$$\phi_D = \phi - \phi_R \tag{5}$$

Thus, with these five equations and the use of a spirule, the phase-angle loci can be quickly constructed. Associated with the constant phase-angle loci are the constant magnitude loci which are nothing more than the set of orthogonal curves. Once the phase-angle loci have been constructed, the magnitude loci can be easily found. A few curves have been drawn with this method to illustrate the general character of the loci, Figs. 2–5. Only the loci in the first quadrant are shown since this is the region of interest in the discussion to follow.

Calculation of Gain and Phase Margins Using Constant Magnitude and Constant Angle Loci

It is now interesting to note that only the -180-degree phase-angle locus and the zero db magnitude locus are needed to be drawn in order to solve the problem. Furthermore, the zero db magnitude locus need only be sketched near the $j\omega$ axis as perpendicular to the -180-degree phase-angle locus. At this point the description can best proceed by the consideration of a specific example. Consider the case of a pair of complex poles at $-4+j5$, $-4-j5$ and a pole at the origin, taking $K=250$. See Fig. 6.

The transfer function is then

$$T(s) = \frac{250}{s(s-4+j5)(s-4-j5)}$$

The point ω_{180} can now be determined by inspection as the intersection of the -180-degree phase locus and the $j\omega$-axis. Scaling the lengths of the various vectors $|T(j\omega_{180})|$ can be calculated as

Table I. Accuracy of Results

Factor	Graphical Approximation	Actual Value	Per-Cent Error
arg $T(j\omega_{180})$	-180 degrees	-178.7 degrees	1.4
$\|T(j\omega_1)\|$	1	1.025	2.5
arg $T(j\omega_1)$	-163 degrees	-166 degrees	1.8

---

* Spirule Company, 9728 El Venado, Whittier, Calif.

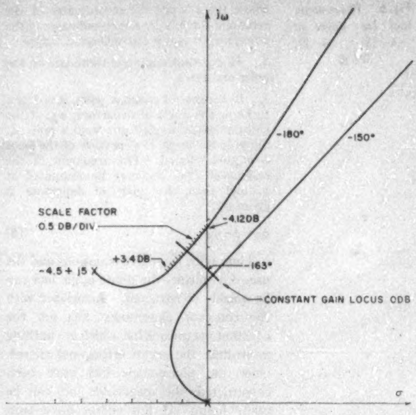


Fig. 7. Phase-angle loci for poles at $-4.5+j5$, $-4.5-j5$, $0+j0$

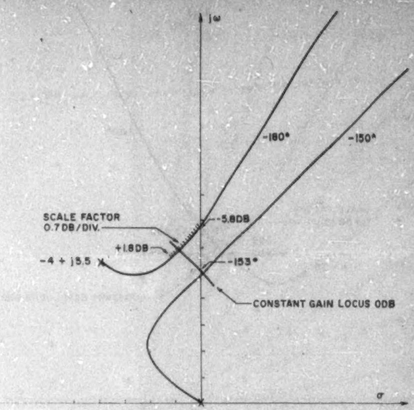


Fig. 8. Phase-angle loci for poles at $-4+j5.5$, $-4-j5.5$, $0+j0$

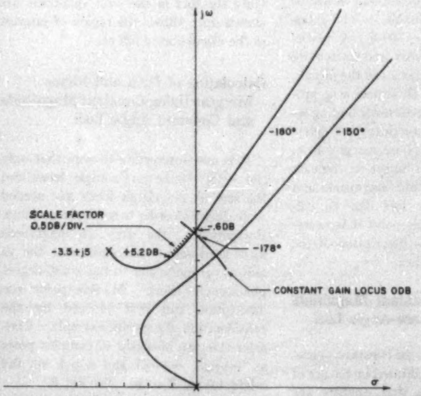


Fig. 9. Phase-angle loci for poles at $-3.5+j5$, $-3.5-j5$, $0+j0$

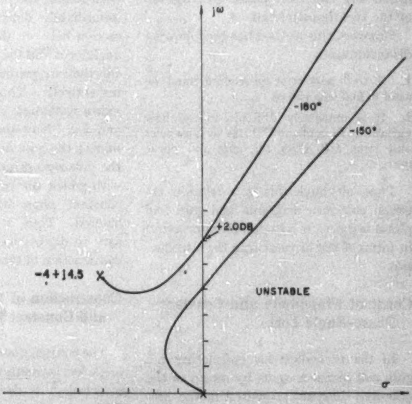


Fig. 10. Phase-angle loci for poles at $-4+j4.5$, $-4-j5.4$, $0+j0$

$$|T(j\omega_{180})| = \frac{250}{(6.3)(12)(4.2)} = 0.78$$

and expressed in db as

$$20 \log_{10} |T(j\omega_{180})| = -20\,(0.11) = -2.2 \text{ db}$$

The next step is to determine the constant gain locus for $T(s)=1$, which may be accomplished in the following manner. Calculate $T(s)$ for one other point along the -180-degree locus, say at a, and express in terms of db.

$$|T(a)| = \frac{200}{(11.1)(5.0)(3.4)} = 1.17$$
$$|T(a)| = 1.4 \text{ db}$$

Now, since distances along the phase-angle loci are very nearly linearly related to $\ln|T(s)|$ a scale factor can be determined on the -180-degree locus, enabling the point for which $|T(s)|=1$ or $|T(s)|=$ zero db to be found. This is accomplished as shown in Fig. 6. The zero db constant gain locus is now sketched as perpendicular to the -180-degree phase-angle locus, and the intersection with the $j\omega$-axis, ω_1, noted. The phase angle at this point may be calculated or found by use of the spirule. This is done in Fig. 6 where $\arg T(j\omega_1) = -163$ degrees and for this example,

GM = 0.22
PM = 17 degrees

Thus, the gain and phase margins have been found. The accuracy of the results will depend on the care with which the loci are drawn, but the values obtained can be quickly checked by direct calculation. This has been done for this example, with the results tabulated in Table I.

If one desires to improve the accuracy of ω_1, a scale factor may easily be determined on the -163-degree locus by using the actual value for $|T(j\omega_1)|$ and calculating $|T(s)|$ for one other point along the locus. The previous procedure may then be used to find a new ω_1, which will necessarily be a more accurate value.

Conclusions

In the previous section a method for calculting the gain and phase margins knowing the positions of the poles and zeros was shown. The method is simple and straightforward, and thus fulfills the first of the requirements.

The second requirement, that the changes in gain and phase margins due to changes in system parameters should be at least qualitatively placed in evidence, is also met. The manner in which changes in the gain factor, **K**, affect the gain and phase margins can be easily visualized. Referring to Fig. 6, a change in **K** would only affect the constant gain locus and would shift a point of constant gain, say the zero db locus, up or down the phase-angle locus. Thus, for a change in **K**, the scale on the −180-degree locus could be easily relabeled and the new gain and phase margins read.

The manner in which changes in the positions of the poles and zeros affect the gain and phase margins can also be visualized. For example, in Fig. 6, a small variation in the position of the complex poles will not affect the slope of the gain and phase-angle loci markedly, and to a very rough approximation, the locus can be considered to move with the pole. In this manner it is possible to tell if the gain and phase margins increase, decrease, or remain the same with specified movement of the singularity. Figs. 7–10 show the phase-angle loci for small variations in the configuration of Fig. 6, and one can become quite adept at visualizing the changes that take place in the loci as the singularities are moved by studying these figures. The −150-degree locus is included for illustrative purposes.

Thus, the construction of the −180-degree phase-angle locus permits a rapid determination of the gain and phase margin of a system from the pole-zero diagram of the open-loop transfer function.

References

1. THE EFFECT OF POLE AND ZERO LOCATIONS ON THE TRANSIENT RESPONSE OF LINEAR DYNAMIC SYSTEMS, J. H. Mulligan, Jr. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 37, 1949, pp. 516–29.

2. SYNTHESIS OF FEEDBACK CONTROL SYSTEMS BY PHASE-ANGLE LOCI, Y. Chu. *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 71, Nov. 1952, pp. 330–39.

3. AUTOMATIC FEEDBACK CONTROL SYSTEM SYNTHESIS (book), J. G. Truxal. McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

Movement of Air in the Electric Wind of the Corona Discharge

MYRON ROBINSON
NONMEMBER AIEE

Synopsis: The mechanism of gas movement in the electric wind is considered herein, and an approximate theory presented which relates relevant electrical and aerodynamic quantities. Among others, these relationships are shown to hold for an electrostatic blower, operating on the electric-wind principle: Gas velocity is a linear function of voltage and is proportional to the square root of current; if the density of the gas is not too low, the efficiency of electrokinetic conversion near sparkover is proportional to the square root of the density, and the velocity at constant current is independent of the density; near sparkover efficiency is independent of voltage; velocity increases slowly as additional blowers are stacked in series; the ozone concentration of blower air resulting from the corona discharge is an increasing function of electric-wind velocity. The forms of the equations relating these variables are found to hold in a variety of cases, even though assumed boundary conditions are not observed experimentally. The practical utility of an electrostatic blower is limited by an efficiency of operation in the neighborhood of 1%.

THE PHENOMENON variously known as the electric wind, corona wind, and electric aura refers to the movement of gas induced by the repulsion of ions from the vicinity of a high-voltage discharge electrode. The effect is most commonly observed with atmospheric air, but other gases and insulating liquids may also be employed as the ion source and carrier medium.

This study will be concerned with the subject of air, and will examine the practicability of using the electric wind as an air-moving mechanism. An electrostatic blower, operating on this principle, directly converts electric energy into the kinetic energy of a moving gas stream. Thus, lacking moving parts, it can have advantages over conventional fans in areas of application where high-voltage low-current power is available; where simple sturdy construction is required; and where noise, gyroscopic effects, and other consequences of rotational motion cannot be tolerated. On the other hand, for gases but not necessarily for liquids, the electrostatic blower is seriously handicapped by an efficiency of electrokinetic conversion of about 1% as compared with 60 or 70% commonly found in conventional fans.

Discovery and Study of Electric Wind

The existence of the electric wind has been known for a long time. As early as 1709, Hauksbee reported experiencing a weak blowing sensation on holding a charged tube to his face.[1] The electric wind rapidly became a favorite experimental device among subsequent investigators, attracting the attention of Newton, Faraday, Maxwell, and a host of others.[2–4]

A realistic quantitative analysis of the wind mechanism was first offered in 1899 by Chattock, who developed a relationship between electric-wind pressure and current for a system of plane-parallel electrodes.[5] Chattock's pressure-current equation was extended to other geometries by Löb in 1954.[6] Three years later, Harney examined the effects on the electrical parameters of a corona discharge which are brought about by the motion of the gas stream.[7] Such effects are small in gases moving at electric-wind velocities, but need not remain so if a dielectric liquid is substituted for the gaseous medium. Paying particular attention to such a case, Stuetzer recently expanded the work of Löb and Harney, giving the most complete account of ion-drag phenomena that has yet been offered.[8–10] A fuller historical review and bibliography appear elsewhere.[11]

Paper 61-90, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 31, 1960; made available for printing December 29, 1960.

MYRON ROBINSON is with Research-Cottrell, Inc., Bound Brook, N. J.

The author is grateful to John Jarema, who made many helpful suggestions, constructed the experimental equipment, and took numerous measurements, and to Robert Brown and George Robertson, who assisted in the ozone determinations. This work was supported by the Bureau of Ships, Department of the Navy, under contract *NObs 77164*.

Explanatory Notes

The mks (meter-kilogram-second) system of units is used throughout this paper. These abbreviations occur: m (meter), kg (kilogram), sec (second), coul (coulomb), v (volt), amp (ampere), w (watt), atm (atmosphere), and pu (per unit). The oblique line /, meaning per, is used with some abbreviations. A dekapoise is equal to 10 poise. All nablas and all boldfaced letters are vectors.

Theory

Electrical Fundamentals

If an electric field is impressed upon an ionized gas, the ions will traverse the field and, by repeated collision with the uncharged molecules present, induce an over-all motion in the surrounding volume of gas. The corona discharge in air offers the simplest means for establishing these conditions.

An electrostatic blower consists essentially of an arrangement of emitting and collecting electrodes, permeable to the gas being moved. Each infinitesimal volume of gas between the electrodes contains an electric charge of density ρ coul/m^3 and is acted upon by an electric field of intensity $\mathbf{E}$ newton/coul (= v/m). The product $\rho\mathbf{E}$, therefore, gives the electrical force exerted on a unit volume of charged gas. Force pu volume is the force pu area pu distance, perpendicular to the area in question, and may be represented as a pressure gradient

$$\nabla p = \rho\mathbf{E} \tag{1}$$

where p is the pressure in newton/m^2. Field intensity and space-charge density are related by Poisson's equation

$$\nabla\cdot\mathbf{E} = \rho/\epsilon \tag{2}$$

where ϵ is the absolute dielectric constant of the gas in farad/m. Equation 2 makes possible the elimination of ρ from equation 1

$$\nabla p = \epsilon\mathbf{E}(\nabla\cdot\mathbf{E}) \tag{3}$$

The impressed voltage V(v) is equal to the difference in potential across the two electrodes. The potential ϕ (v) anywhere in the system is given by

$$\mathbf{E} = -\nabla\phi \tag{4}$$

The passage of current between the electrodes is the result of two simultaneous effects: (1) movement of charge relative to the main body of gas and (2) transport of charge by, rather than through, the gas stream. Assuming that the space charge consists wholly of ions of one kind, a condition that is closely satisfied outside the active region of corona glow, the first component of velocity is $b\mathbf{E}$, where b is the ionic mobility in (m/sec)/(v/m). If $\mathbf{v}$ is the gas velocity (m/sec) relative to the electrode system, i.e., the velocity of the electric wind, the total ionic velocity $\mathbf{v}_t$ is

$$\mathbf{v}_t = \mathbf{v} + b\mathbf{E} \tag{5}$$

The current density $\mathbf{j}$ (amp/m^2) in the region between the electrodes is then given by

$$\mathbf{j} = \rho\mathbf{v}_t = \rho(\mathbf{v} + b\mathbf{E}) \tag{6}$$

Eliminating ρ between equation 2 and equation 6

$$\mathbf{j} = \epsilon(\nabla\cdot\mathbf{E})\,(\mathbf{v} + b\mathbf{E}) \tag{7}$$

From equations 3 and 4, the generated pressure can be expressed as a function of the applied voltage. Equations 7 and 4 give the current in terms of the voltage. A relationship linking pressure and velocity is required, and the performance of the electrostatic blower will be completely described.

Electrokinetic Considerations

Air will move through a blower at the velocity at which the forward electrical pressure is matched by the aerodynamic back pressure. The latter quantity may be expressed as

$$p = K\rho_g v^2/2 \tag{8}$$

where K, an appropriate loss coefficient, is a function of geometry and, in many cases, almost independent of velocity.[11–13]

Equations 1 and 6 yield another useful relationship

$$\nabla p = (\mathbf{j} - \rho\mathbf{v})/b \tag{9}$$

The quantity $\rho\mathbf{v}$ is the extent to which the current density is increased by the carrier fluid moving under the influence of migrating ions. In liquids, $\rho\mathbf{v}$ may be comparable to $\mathbf{j}$; in gases it is, for present purposes, negligible.

Then

$$\Delta p \approx \mathbf{j}/b \tag{10}$$

Attention will be confined to plane-parallel, concentric-spherical, and coaxial-cylindrical electrode configurations. A plane electrode supporting a corona discharge is, of course, not realizable in practice. A plane-parallel system is, nevertheless, an adequate approximation to certain practical geometries. Concentric-spherical electrodes, in their simplest experimental form, comprise a point discharge and a segment of a spherical surface, the latter serving as the collecting electrode. Coaxial cylinders are conveniently represented by a discharging wire and a segment of a coaxial-cylindrical surface.

It is assumed that the electrical and aerodynamic boundary conditions are identical. Hence, the directions of the vectors $\mathbf{E}$, $\mathbf{j}$, and $\mathbf{v}$ are taken to be the same at any given point in the interelectrode region. In order to conform to this restriction, only the plane-parallel blower can be built in a duct of uniform cross section. The spherical electrodes must be housed in a conical duct and the cylindrical electrodes in a wedge-shaped duct. It will be found, however, that marked departures from these conditions need not invalidate the functional forms of the relationships between the variables.

In each of the three co-ordinate systems mentioned, the pressure gradient is given by

$$|\nabla p| = dp/dr \tag{11}$$

where r is the distance from the discharge electrode. Thus, for gases, the pressure differential maintaining the electric wind is

$$p = \int_{r_d}^{r_c} \frac{j}{b}\,dr \tag{12}$$

where r_d and r_c are, respectively, the coordinates of the discharge and collecting electrodes.

Assuming constant mobility, equations 8 and 12 combine to give

$$v_c = \left[\frac{2}{K\rho_g b}\int_{r_d}^{r_c} j\,dr\right]^{1/2} \tag{13}$$

where $j = j\ (r,\ r_c,\ r_d)$. In an expanding flue, the loss coefficient K is an increasing function of r_c. The quantity v_c designates the velocity of the gas as it passes through the collecting electrode. The current density is dependent on the geometry of the system and is, in addition, proportional to the current. Consequently

$$v_c = g_1\ (i/\rho_g bK)^{1/2} \tag{14}$$

where g_1 is an appropriate function of the geometry (m$^{-1/2}$) and i is the current (amp).

Since in gases, the electric-wind velocity is much less than the ionic velocity relative to the gas, equation 7 simplifies to

$$\mathbf{j} = b\mathbf{E}\nabla\cdot\mathbf{E} \tag{15}$$

From the solution to this equation, the current may be expressed as a function of the voltage. By assuming an appropriate form for the electric-field strength at the discharge electrode, this current–voltage relationship has been shown, for certain geometries, to be

$$i = g_2 \epsilon b V(V - V_s) \tag{16}$$

where $g_2(\mathrm{m}^{-1})$ is a function of the geometry and V_s is the apparent corona-starting voltage. Equation 16 finds experimental confirmation over a much wider range of geometries than that for which analytical solutions of equation 15 are available, and may be assumed generally true for space-charge-limited discharges of the kind considered.[14]

Substituting equation 16 in equation 14 and setting $g_1 g_2^{1/2} = g_3$ (m^{-1}), there results

$$v_c = g_3(\epsilon/\rho_g K)^{1/2}[V(V - V_s)]^{1/2} \tag{17}$$

The main region of interest is $V >> V_s$. By expanding the bracketed term, an approximate linear relationship between velocity and voltage is obtained:

$$v_c = g_3(\epsilon/\rho_g K)^{1/2}\ (V - V/2 + \ldots) \tag{18}$$

The kinetic energy of a mass of air M (kg) emerging from a blower in a time interval t (sec) is $Mv^2/2$. The kinetic power output P_o (w) is then given by

$$P_o = \frac{d}{dt}\left(\frac{1}{2}Mv_c^2\right) = \frac{1}{2}\frac{dM}{dt}v_c^2 = \frac{1}{2}\rho_g A v_c^3 \tag{19}$$

where $A(\mathrm{m}^2)$ is the cross-sectional area of the duct at $r = r_c$. Alternatively, using equations 17 and 19

$$P_o = (g_7^2/\rho_g K^3)^{1/2}\,[V(V - V_s)]^{3/2} \tag{20}$$

where $g_7 = g_3^2 A$ (m^{-1}).

The fractional efficiency of electrokinetic conversion η is defined as

$$\eta = P_o/P_i \tag{21}$$

where $P_i = Vi$ is the electric power input (w).

Setting

$$g_4 = g_1^2 g_2^{1/2} A/2 \text{ (dimensionless)} \tag{22}$$

it follows that

$$\eta = (g_4/b)\,(\epsilon/\rho_g K^3)^{1/2}\,[(V - V_s)/V]^{1/2} \tag{23}$$

To achieve maximum output, a blower must be operated close to sparkover. The efficiency normally attained in practice is, therefore,

$$\eta = \eta_0 = (g_4/b)\,(\epsilon/\rho_g K^3)^{1/2} \tag{24}$$

The quantities b, ϵ, and ρ_g are known. The loss coefficient K may be estimated, once the geometry is selected. But, for greatest accuracy, it is best determined experimentally. The functions g may be computed—subject to certain simplifying, and perhaps invalidating, assumptions. They should, in all cases, be checked experimentally. Sample calculations of the g's have appeared previously.[11]

Pressure Effects

It is shown by experiment and theory that the mobility of a gas over a wide range is

$$b = \rho_{go} b_o/\rho_g \tag{25}$$

where the subscript o designated standard conditions.[15] Rewritting equation 24

$$\eta_0 = (g_4/\rho_{go} b_o)\ (\epsilon,{}^{3}) K^{1/2}\,\rho_g^{1/2} \tag{26}$$

The dielectric constant ϵ is sensibly constant over a wide range of gas densities. The efficiency, therefore, increases with density, but is very low in atmospheric air. Because of this latter condition, special interest will center on explorations into blower operation at other than atmospheric conditions.

Ozone Generation

The glow surrounding the emitting electrode of a corona discharge in air and certain other gases may constitute a region of significant chemical activity. Ozone and, to a lesser extent, oxides of nitrogen are formed in air. These gases are toxic to humans—0.1 ppm (parts per million) of ozone in air is commonly accepted as the maximum allowable concentration for prolonged periods of exposure.[16]

The rate at which ozone is generated in the electric discharge is a function of the current and the electrode geometry. Assume that for a given geometry this rate G(kg/sec) is given by

$$G = Ri^a \tag{27}$$

where R (kg-sec^{a-1}/coul^a) is a constant of proportionality. The resulting mass concentration of ozone C_m (mass of ozone per mass of air) is

$$C_m = Ri^a/v_c A \rho_g \tag{28}$$

The corresponding volume concentration C_v (volume of ozone per volume of air) is

$$C_v = Ri^a/v_c A \rho_z \tag{29}$$

where ρ_z is the density of ozone.

The foregoing concentration formulas are valid for series and parallel combinations of identical blowers. In both cases, i is the total current drawn by the multiblower system, and A is the total cross-sectional area which, in a series arrangement, is unchanged. Eliminating the current between equations 14 and 29 and setting

$$1/A g_1^{2a} = g_5(\mathrm{m}^{a-2})$$

the result is

$$C_v = g_5 R(\rho_g bK)^{a} v^{2a-1}/\rho_z \tag{30}$$

For $a > 0.5$, it is clear that any attempt to dilute the ozone by increasing the airflow rate will be futile.

The difficulty existing when $a > 0.5$ is, to some extent, remediable. Equation 30 reveals that it is possible to reduce the ozone concentration at constant volumetric flow rate by working several individual blowers in parallel, each at reduced capacity. Specifically, a required volumetric flow vA (m^3/sec) can be provided by a single blower of capacity vA, or by n identical parallel blowers, each moving air at the velocity v/n. The parallel assembly has a cumulative cross-sectional area nA. The ozone concentration is the same for the system of blowers as for any individual in it; hence, for the system

$$C^v = g_5 R(\rho_g bK)^a (v/n)^{2a-1}/\rho_z \tag{31}$$

The ozone concentration has now been diminished by a factor of n^{1-2a}.

The efficiency, in terms of the number of parallel units, may be obtained from equations 17 and 23

$$\eta = \frac{g_4}{b}\left(\frac{\epsilon}{\rho K^3}\right)^{1/2}\left\{1 - \frac{2}{1 + \left(1 + \frac{4U^2\rho_g K}{g_6 \epsilon n^2 V_s^2}\right)^{1/2}}\right. \tag{32}$$

where $g_6 = g_3^2 A^2$ and $U = nvA$, the total volumetric flow rate. Reducing the ozone concentration in the manner described will also bring about a decrease in efficiency.

Blowers in Series

If the pressure developed by a single blower is small, or the back pressure of the external load large, it is of particular interest to investigate the advantage of stacking blowers in series. If each blower in the series assembly generates a pressure p, the whole assembly then generates pressure np. This forward pressure must overcome the internal back pressure of each blower p_B

$$p_B = K_B \rho_g v_c^2/2 \tag{33}$$

where K_B is the internal loss coefficient of a single blower. The forward pressure must also accommodate an external load of back pressure

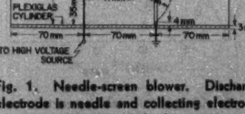


Fig. 1. Needle-screen blower. Discharge electrode is needle and collecting electrode wire screen in form of sphere segment

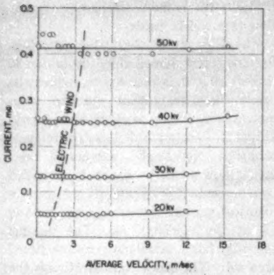


Fig. 2. Effect of air velocity on corona current. Corresponding velocity-current values for electric wind in needle-screen blower are shown by dashed parabola (equation 14)

$$p_L=K_L\rho_g v_c^2/2 \tag{34}$$

where K_L is the loss coefficient of the external load.

Then

$$np=np_B+p_L=nK_B\rho_g v_c^2/2+K_L\rho_g v_c^2/2 \tag{35}$$

or

$$v_c=[2p/\rho_g(K_B+K_L/n)]^{1/2} \tag{36}$$

The total loss coefficient K_B+K_L/n here corresponds to the coefficient K in equation 8. Consequently, by comparison with equation 14

$$v_c=g_1[i/\rho_g b(K_B+K_L/n)]^{1/2} \tag{37}$$

where i is the current fed to a single blower.

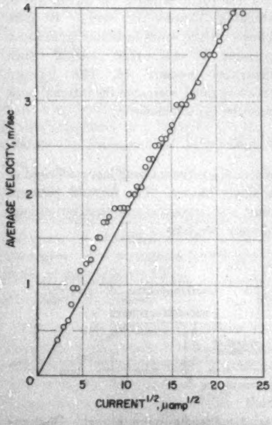


Fig. 3. Characteristic velocity-current relationship. Line is that of equation 14

Experimental Results and Discussion

Blower Design

Most of the experimental data to be reported were obtained, using an electrostatic blower not conforming in design to any of the geometric configurations already described. The discharge electrode in this blower, Fig. 1, consists of a needle point, the tip of which is concentric with a wire screen serving as the collecting electrode. The electrodes are housed in an insulating cylinder.

Unless so stated to the contrary, all measurements hereafter given were obtained with a negative discharge with this blower in atmospheric air. The blower is not of the concentric spherical type described, since the same boundary conditions are not imposed aerodynamically as well as electrically. It is, however, simpler to construct and does give an idea of the extent to which the mathematical assumptions already made can be violated without seriously impairing the validity of the forms of the derived equations. Other experimental designs will be described later.

Effect of Fluid Flow on Current

In the foregoing equations, dependent on equation 7, it is assumed that $\mathbf{v} \ll b\mathbf{E}$. This is equivalent to saying that the current is essentially independent of the gas velocity. The truth of this assumption is examined in Fig. 2. Current is given as a function of average air velocity. Along the broken curve, air motion is produced solely by the mechanism of the electric wind. An external fan provides the pressure required to reach other velocities. The current is constant over a velocity range extending well beyond the maximum attained by the electric wind.

Velocity

Equation 14 predicts that the velocity of the electric wind will be proportional to the square root of the current. This expectation is supported by the experimental data; see Fig. 3.

The term $g_1/K^{1/2}$ is obtained from the slope of the line of Fig. 3. By combining this with the values of g_2 and V_s, established from current-voltage data, any of the quantities current, voltage, velocity, power, and efficiency can be calculated in terms of the others. In principle $g_1/K^{1/2}$, g_2, and V_s can be determined for a given blower simply by measuring the current and velocity corresponding to a single arbitrary voltage, $V>V_s$, not too close to sparkover.

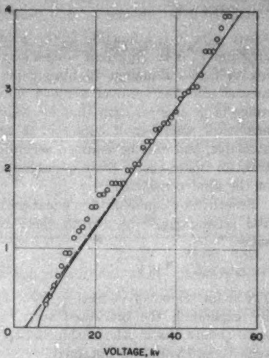


Fig. 4. Electric-wind velocity as function of voltage. Solid curve conforms to equation 17; broken line to equation 18

Air velocity as a function of voltage is shown in Fig. 4. The solid line represents equation 17, using the value $g_1/K^{1/2}$ already known from experiments. It is seen that the curve is almost linear, closely following equation 18.

Efficiency

In most practical applications, the efficiency of electrokinetic conversion is of primary concern. For a set geometry, equation 24 reveals that the efficiency is proportional to $\epsilon^{1/2}/b\rho_g^{1/2}$. In a negative discharge in atmospheric air, this quantity is equal to 0.014, but for liquids it is much larger; e.g., for kerosene it is about 0.5. In general, the efficiency of a liquid pump will exceed that of the corresponding gas pump by 1 or 2 orders of magnitude. In view of this, low gas blower efficiencies should be anticipated.

Efficiencies calculated on a purely theoretical basis are of interest for comparative purposes. The results charted in Table I were obtained from equation 24 for parallel-plate blowers in room air. The loss coefficient K varies with duct design, but is in no case less than unity. It is easily shown that g_4 is in the neighborhood of 1.5 or 2.[11]

Table I. Efficiencies for Parallel-Plate Blowers

K	η, %			
	−Ions		+Ions	
	$g_4=2$	$g_4=3/2$	$g_4=2$	$g_4=3/2$
1	2.7	2.0	3.0	2.6
2	1.0	0.7	1.4	0.9
3	0.5	0.4	0.7	0.5

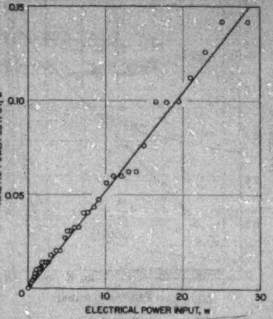


Fig. 5. Variation of kinetic power output of moving air column with electric power input. Curve follows theoretical relationship

Blowers of this type, it seems, are inherently limited to efficiencies of the order of 1%. Cylindrical or spherical geometries do not materially affect this disappointingly low value.

Fig. 5 gives the experimental data and theoretical curve (equations 21 and 23), comparing output and input powers in the needle-screen blower under consideration. The changing slope indicated in equation 23 is not evident in the curve. Although efficiency is sensitive to voltage changes at low voltage, low voltage yields such low power that the expected dip in the curve occurs too close to the origin to be noticed.

Plotting efficiency as a function of voltage in Fig. 6, the sought-after effect becomes apparent. The curve is that of equation 23. The sharp deviation from theory in the neighborhood of 20 kv is easily explained. Since efficiency is proportional to v^3, relatively small instrumental errors in measuring velocity are grossly magnified in efficiency calculations. Note the experimental velocity readings between 15 and 25 kv in Figs. 3 and 4.

PRESSURE AND DENSITY EFFECTS

Since the product $\rho_g b$ is independent of the pressure, the gas velocity as given by equation 14 likewise must be independent. To confirm this behavior, the velocity output of a needle-screen blower was measured in air at room temperature, over a range of ambient pressures from a fraction of an atmosphere to 8 atm. The current was held constant at several pressure levels by adjusting the voltage as the pressure varied. The results are reproduced in Fig. 7. For air $\rho_g b$ is 2.4×10^{-4} coul-sec/m^3. Fig. 3 gives $g_1/K^{1/3}=2.79$ m$^{-1/3}$. Using these values, equation 14 yields $v_e=180\ i^{1/3}$, a relationship closely followed by the data.

As the pressure is raised, the ion mobility decreases, causing the current at a given voltage to fall. In order to maintain a constant current, the applied voltage must be raised with the pressure. Experiments reveal that although the sparkover voltage increases, with the pressure, it does not rise fast enough to permit a constant current to be maintained without exceeding sparkover. Consequently, the lines of constant velocity are bounded at high pressures as shown. With decreasing pressure, the sparkover voltage drops more rapidly than does the applied voltage required to sustain the constant current. Lines of constant velocity are, therefore, bounded by sparkover on the low-pressure side, also.

Equation 26 indicates that efficiency increases with gas density. It is important to note, in this connection, that mere pressurization, without regard to temperature, may not be sufficient to raise the efficiency. It is not the pressure per se, but the density which is the governing factor.

The product $\rho_g b$ has an approximately constant value for a given gas; i.e., it is independent of density, ambient pressure, and temperature. It follows at once from equation 14 that the velocity of the electric wind is likewise independent of these three quantities. Changes in efficiency brought about by manipulating pressure or temperature result from altering the density and, with it, the mass of gas moved pu time.

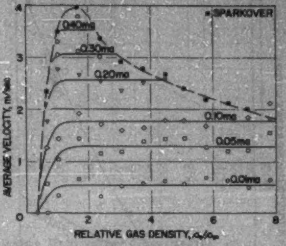


Fig. 7. Electric-wind velocity as function of air density and corona current. Straight-line portions of the curves closely agree with equation 14

According to equation 14, the velocity is inversely proportional to $(\rho_g b)^{1/3}$. Fig. 8 illustrates the extent to which this quantity varies among a number of gases. Ammonia and very pure argon are separated by 3 orders of magnitude. The reduction in mobility caused by impurities and by molecules of electronegative gases is well-known.[16]

In order to achieve maximum output power, an electrostatic blower must be operated at maximum input power; i.e., at voltages just below sparkover. Operation during sparkover results in lowered gas velocity, probably because sparking not only produces mobile free electrons, but also confines the current to an extremely small region of the total volume between the electrodes.

Increasing the density of a gas will, except at sufficiently low densities, raise the sparkover voltage. Operating close to sparkover gives $V_{sp}>>V_s$, where the subscript *sp* designates the maximum presparkover voltage. Equations 17 and

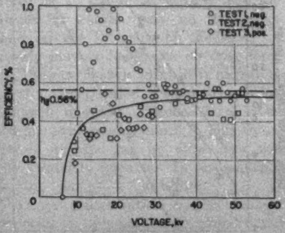


Fig. 6. Example of variation of efficiency with voltage. Curve is that of equation 23

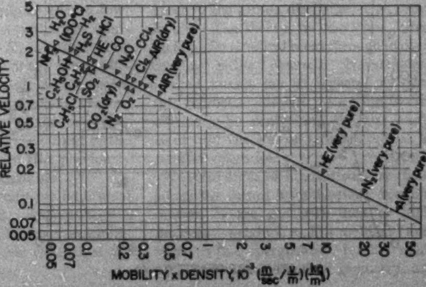


Fig. 8. Electric-wind velocities of various gases relative to dry air. Comparisons made at constant negative current

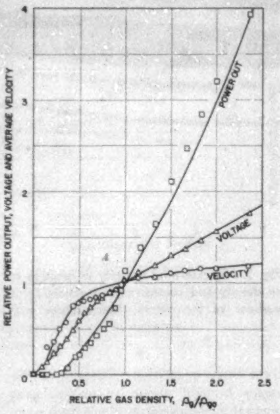


Fig. 9. Power output, voltage, and velocity just below sparkover in terms of gas density. All measurements relative to those at atmospheric conditions. Power and velocity curves calculated from equations 20 and 17 respectively

20 then yield the convenient approximations $v_{esp} \alpha V_{sp}$ and $P_{esp} \alpha V_{sp}^2$.

Experimental velocity, power, and voltage data just below sparkover are shown as functions of gas density in Fig. 9. The smooth curves of velocity and power represent equations 17 and 23 respectively. Efficiency at sparkover is, by equation 24, proportional to $\rho_g^{1/2}$, which relationship is borne out by Fig. 10.

The quantities g_1 through g_8 are, by definition, functions of geometry only. Accordingly, they must remain fixed for a given electrode arrangement, regardless of electrical conditions or the nature of the gas. In considering the effects of pressure, certain assumptions have been made concerning the variation of mobility with density. At low enough densities, however, for reasons given elsewhere, these assumptions fail.[15] It will now be determined at what point and to what extent this occurs.

All the functions g are defined in terms of g_1, g_2, and A. It will therefore, suffice to examine g_1 and g_2 alone. To do this, no limits are placed on the validity of equation 25, and then the apparent changes in g_1 and g_2 are observed as the density varies. The experimental results calculated from equations 14 and 16 for a needle-screen blower in air appear in Fig. 11. The quotient $g_1/K^{1/2}$ rather than g_1 alone is given since the former is more readily measured. Both $g_1/K^{1/2}$ and g_2 are constant at densities down to about atmospheric level, but below this, $g_1/K^{1/2}$ falls to zero and g_2 increases rapidly. These variations result from the assumption that b remains proportional to $1/\rho_g$ at low ρ_g. In reality, b rises faster in this region than does $1/\rho_g$, and this disproportionate increase is, by equation 11, reflected as a decrease in $g_1/K^{1/2}$. Similarly, by equation 16, g_2 must rise.

The inconstancy of the g's below a certain density does not affect the usefulness of the general theory below that density. At low densities, equation 25 and the relations containing it should be disregarded and appropriate values of mobility be introduced directly into the theory and not as functions of density.

OZONE GENERATION

Equations 28, 29, and 30 assume that the rate of ozone generation for a given geometry is proportional to a power of the current. Experimental support for this hypothesis is offered in Fig. 12. Ozone concentration was determined by the potassium-iodide method of Britt.[17]

Electrostatic precipitators designed to filter air for human consumption commonly employ a positive discharge, since reason exists for believing that such polarity will yield a lower ozone output.[18,19] Similar considerations evidently do not apply to the needle-screen blower. Contrary to expectation, the curves indicate that positive corona generates ozone at several times the rate of negative corona. The negative curve for a single blower is fitted by this rewriting of equation 27:

$$G = 0.1\, i^2 \tag{38}$$

Day-to-day variations in the data are, perhaps, attributable to fluctuations in atmospheric conditions.[20]

A blower, operating with negative corona near sparkover, generates about 0.8 ppm by volume of ozone in air, a quantity in excess of the usual acceptable limit. This ozone concentration may be reduced at constant volumetric flow rate by adopting the arrangement of parallel blowers previously described. Experimental and calculated data for such a case are shown in Fig. 13. Ozone concentration measurements were actually made with a single blower over a range

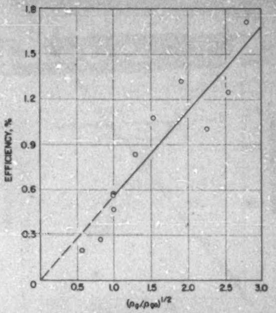


Fig. 10. Efficiency plotted against relative air density. Straight line conforms to equation 26

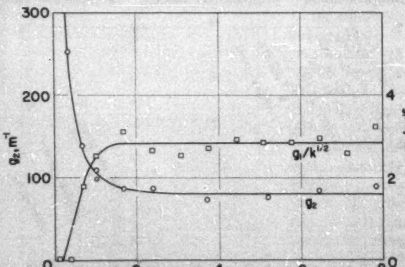
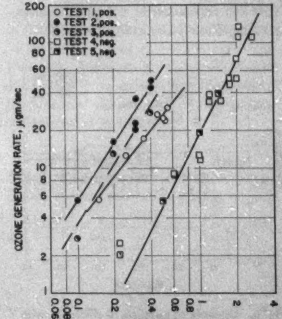


Fig. 11 (left). Range of constancy of the geometric functions $g_1/K^{1/2}$ and g_2

Fig. 12 (right). Examples of ozone-generation rate in terms of corona current. Data for five blowers in series



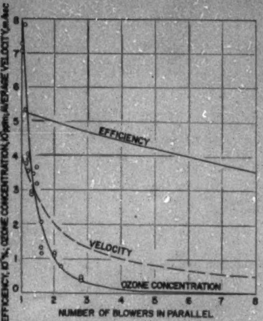


Fig. 13. Efficiency, ozone concentration by volume, and velocity resulting from running different numbers of blowers in parallel at constant total volumetric flow rate. Efficiency and ozone curves drawn from equations 32 and 31 respectively

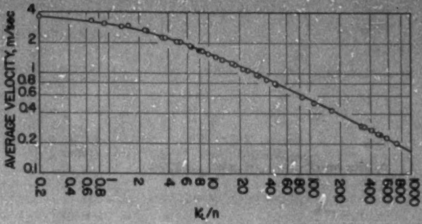


Fig. 14. Effect of external load on air velocity. Curve follows equation 36

of currents; for this reason, experimental points corresponding to fractional blowers appear. In Fig. 13, the assumed parallel system is taken to have a constant volumetric flow rate of 1.54×10^{-3} m^3/sec, the maximum delivered by a single blower. The drop in ozone concentration, as blowers are added in parallel, is striking. Two blowers produce 13% of the concentration of a single blower and five blowers only 0.8% of that amount. The efficiency loss accompanying a large decrease in ozone concentration is small.

POLARITY

Performance can be expected to be generally dependent on the choice of polarity of the discharge when the ionic mobility appears in the governing equations. In atmospheric air, the mobility of stable negative ions is 2.0×10^{-4} (m/sec)/(v/m) and that of stable positive ions 1.4×10^{-4}. Inserting these values into equations 13 and 23, $v_e^+ \approx 1.1\, v_e^-$ at constant current, and $\eta^+ \approx 1.4\, \eta^-$. The force on an ion in an electric field depends, of course, only on its charge and on the intensity of the field. A slowly moving ion exerts its drag force on the gas through which it is migrating over a longer period of time than an ion of high mobility. Hence, at a given current, the momentum $(\int F dt)$ transferred to the gas by slow ions will be relatively greater. In consequence $v_e^+ > v_e^-$ and $\eta^+ > \eta^-$, as before.

Ozone concentration C_v at constant velocity is, by equation 30, proportional to Rb^a. In the needle-screen blower, a reversal of polarity will cause R and b to change in opposite directions. Variation, if any, of the exponent a in Fig. 12 is not clear. Nevertheless, if minimum C_v is desired at a given output, negative corona remains preferable.

According to equation 17, velocity will be independent of polarity for a given voltage if the small effect of the corona starting voltage is neglected. This has been shown experimentally. Sparkover during negative operation occurs in this case at about 1.8 times the positive sparkover voltage. Since the kinetic power delivered at sparkover is approximately proportional to the cube of the sparkover voltage, a single negative blower can produce the power output of six positive units.

BLOWERS IN SERIES

Equation 37 predicts the velocity to be expected from a series arrangement of blowers working against a load. For simplicity, the loss coefficients will continue to be considered independent of velocity. Screen coefficients are, however, known to increase at low enough velocities.[13] At normal operating voltages, such low velocities result from large external loads. In that case, K_L/n is likely to be so much larger than K_B that departures from "constant" value of the latter do not materially affect the velocity calculated from equation 37.

Fig. 14 shows the velocity output of five blowers in series as a function of the external load per blower. Since the blower assembly is designed with a low-loss inlet and re-entrant outlet, the minimum possible value of K_L is unity. The characteristic g_1 of each of the five needle-screen blowers employed in the series test is, for an undetermined reason, significantly lower than in the preceding cases. Because of this, the output of the assembly does not exceed that previously obtained for an individual unit.

The increase in velocity per additional series blower is small. Setting $K_L = 1$, Fig. 14 shows that a 5-blower assembly delivers 3.5 m/sec of air as against 3.0 m/sec for a single blower.

OTHER ELECTRODE GEOMETRIES

The described experimental results were obtained by using the needle-curved screen blower shown in Fig. 1. Many of the experiments were, however, repeated for electrode configurations of other sizes and designs; namely: wires and plane screens, wires and rods, points and plane screens, points and rings. Except for the anticipated variations in the constants g and K, the results proved similar to those already given.

Conclusions

An approximate theory has been developed relating the electrical and mechanical parameters of an electric-wind blower. Although derived for specific boundary conditions, the forms of the equations are found to be widely applicable and should be useful in predicting the performance of electrostatic blowers of a number of diverse geometries, provided only that two geometric constants are determined experimentally.

Complete lack of moving parts gives the electrostatic blower these advantageous characteristics over the conventional fan: vibrationless, almost silent operation; absence of gyroscopic and other rotational effects; no lubrication requirements; and minimal replacement of worn parts (discharge electrodes). The chief disadvantages of the blower are the extremely low efficiency of energy conversion and, in certain cases, the generation of extraneous gases in the corona discharge; e.g., ozone in air. The theory, as developed, offers no hope for a radical breakthrough in overcoming either obstacle. Nevertheless, limited success can be expected by careful geometric design; e.g., improving efficiency by controlling the loss coefficient and the g's, or reducing ozone by parallel operation.

Several areas of interest suggest themselves for future investigations: contamination of the gas with dust or high-molecular-weight (i.e., low-mobility) ions, pulse or a-c energization, further study of the g's as functions of geometry, and combination of the blower with an electric precipitator.

Nomenclature

a = constant defined by equation 27, dimensionless
A = cross-sectional area, m^2
b = ion mobility, (m/sec)/(v/m)
b_0 = ion mobility under standard conditions, (m/sec)/(v/m)
C_m = mass concentration of ozone, dimensionless
C_v = volume concentration of ozone, kg/m^3
E = electric field strength, v/m
g_1 = function of geometry, m$^{-1/2}$
g_2 = function of geometry, m^{-1}
g_3 = function of geometry, m^{-1}
g_4 = function of geometry, dimensionless
g_5 = function of geometry, m^{-1}
g_6 = function of geometry, m^2
g_7 = function of geometry, m^{-1}
G = time rate of ozone generation, kg/sec
i = current, amp
j = current density, amp/m^2
K = aerodynamic loss coefficient, dimensionless
K_B = internal loss coefficient of a single blower, dimensionless
M = mass, kg
n = number of blowers in series or parallel, dimensionless
p = pressure, newton/m^2
p_B = internal back pressure of a single blower newton/m^2
P_i = electrical power input, w
P_o = kinetic power output, w
P_{osp} = maximum presparkover kinetic power output, w
r = distance from origin of co-ordinates, m
r_c = co-ordinate of collecting electrode, m
r_d = co-ordinate of discharge electrode, m
R = rate of ozone generation, kg/coul
t = time, sec
U = total volumetric flow rate of parallel assembly, m^3/sec
v = wind velocity, m/sec
v_c = wind velocity at collecting electrode, m/sec
v_{csp} = maximum presparkover wind velocity at collecting electrode, m/sec
v_i = ion velocity, m/sec
V = voltage, v
V_s = apparent corona starting voltage, v
V_{sp} = maximum presparkover voltage, v
ϵ = absolute dielectric constant, farad/m
η = fractional efficiency of electrokinetic conversion, dimensionless
η_0 = asymptotic fractional efficiency of electrokinetic conversion, dimensionless
ρ = charge density, coul/m^3
ρ_g = gas density, kg/m^3
ρ_{g0} = gas density at standard conditions, kg/m^3
ρ_z = ozone density, kg/m^3
ϕ = electric potential, v

References

1. Physico-Mechanical Experiments on Various Subjects (book), **Francis Hauksbee.** London, England, 1709, pp. 46–47.

2. Opticks (book), **Isaac Newton.** London, England, 1718, pp. 315–16.

3. Experimental Researches in Electricity (book), **Michael Faraday.** London, England, vol. 1, 1839, pars. 665, 1443, 1444, 1535, 1592, 1595.

4. Treatise on Electricity and Magnetism (book), **James Clerk Maxwell.** Oxford, England, vol. 1, 1873, pp. 52–54.

5. On the Velocity and Mass of Ions in the Electric Wind in Air, **A. P. Chattock.** *Philosophical Magazine*, London, England, vol. 48, 1899, pp. 401–20.

6. The Pressure Effects of Ion Currents in Atmospheric Air for Different Electrode Configurations, **E. Löb.** *Archiv der Elektrischen Übertragung*, Berlin, Germany, vol. 8, 1954, pp. 85–90.

7. An Aerodynamic Study of the Electric Wind, **D. J. Harney.** *M. S. Thesis*, California Institute of Technology, Pasadena, Calif., 1957; also *ASTIA Document AD-134400*, Armed Services Technical Information Agency, Washington, D. C., 1957.

8. Ion Drag Pressure Generation, **O. M. Stuetzer.** *Journal of Applied Physics*, New York, N. Y., vol. 30, 1959, pp. 984–94.

9. Ion Drag Pumps, **O. M. Stuetzer.** *Ibid.*, vol. 31, 1960, pp. 136–46.

10. Instability of Certain Electrohydrodynamic Systems, **O. M. Stuetzer.** *Physics of Fluids*, New York, N. Y., vol. 2, 1959, pp. 642–48.

11. Movement of Air in the Electric Wind of the Corona Discharge, **Myron Robinson.** *ASTIA Contract NObs 77164*, Armed Services Technical Information Agency, June 8, 1960.

12. Woods' Practical Guide to Fan Engineering (book), **W. C. Osborne, C. G. Turner.** Woods, Colchester, England, 1954, pp. 80 ff.

13. Changes Produced in an Air Stream by Wire Gauze, **P. Jonas.** *Engineers' Digest*, London, England, vol. 18, 1957, pp. 191–93.

14. Handbook of Physics, **L. B. Loeb** (S. Flügge, editor). Springer Verlag, Berlin, Germany, 1956, vol. 22, p. 513.

15. Fundamental Processes of Electrical Discharges in Gases (book), **L. B. Loeb.** John Wiley & Sons, Inc., New York, N. Y., 1939, chap. 1.

16. Bibliography of Ozone Technology (book), **C. E. Thorp.** Armour Research Foundation, Chicago, Ill., 1955, vol. 2, pp. 128–34.

17. Ozone Measurements, **J. R. Britt.** *NRL Report C-3475*, Naval Research Laboratory, Washington, D. C., pp. 2–4.

18. Electrochemistry of Gases and Other Dielectrics (book), **G. Glocker, C. S. Lind.** John Wiley & Sons, Inc., 1939, chap. 2.

19. Particle Charging in Electrostatic Precipitation, **H. J. White.** *AIEE Transactions*, vol. 70, pt. II, 1951, pp. 1186–91.

20. Influence of Water Vapor on Ozonizer Efficiency, **C. E. Thorp, W. J. Armstrong.** *Industrial and Engineering Chemistry*, Washington, D. C., vol. 38, 1946, pp. 1314–21.

A Low-Level Linear Rundown Circuit for Pulse-Height to Pulse-Width Conversion

DAVID L. ENDSLEY
NONMEMBER AIEE

W. W. GRANNEMANN
NONMEMBER AIEE

THOMAS SUMMERS
NONMEMBER AIEE

Synopsis: The two-transistor linear rundown circuit analyzed here is desirable because of its size and economy. The circuit uses one transistor for the constant-current generator and another for an emitter-follower. Waveforms obtained from both experimental and theoretical analysis correlated well enough to plot them as a single waveform. With the addition of a differentiating circuit, this rundown circuit is well suited for the problem of low-level pulse-height to pulse-width conversion.

MOST available methods for converting pulse height to pulse width use high-amplitude pulses requiring linear amplifiers, but its size and economy make the low-level two-transistor circuit used here most desirable. It achieves linear rundown, and a conversion from pulse height to pulse width can be achieved by adding a differentiating circuit. The circuit forms a linear rundown activated by a constant-current generator discharging a condenser. Because the rundown length is directly proportional to pulse height, it can convert pulse height to width efficiently.

Circuit Analysis

The rundown circuit, Fig. 1, uses one transistor for the constant-current generator and another for the emitter-follower. Fig. 2 shows the small signal equivalent circuit for Fig. 1; its analysis is in two parts because the values for C_e and r_e differ for intervals during and after pulse application. During pulse (a), transistor T_1 is biased in the forward direction. The emitter resistance, r_e, is 40 ohms; C_e is primarily diffusion capacity.[1] Diffusion capacity is expressed by $C_e = 1/1.5\omega_\alpha r_e$, where ω_α is transistor cutoff frequency and r_e its emitter resistance.

Paper 61-1, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted February 23, 1960; made available for printing October 5, 1960.

David L. Endsley, W. W. Grannemann, and Thomas Summers are with the University of New Mexico, Albuquerque, N. Mex.

The authors wish to express their thanks to Robert Tinnell for computing some of the curves.

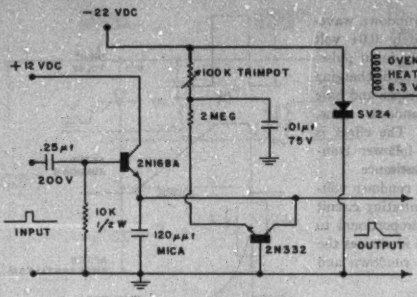


Fig. 1. Rundown circuit

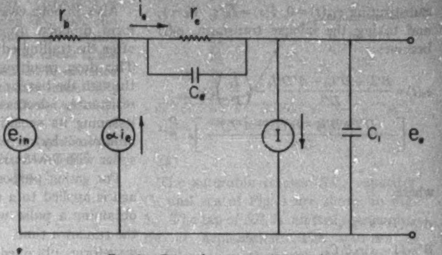


Fig. 2. Equivalent circuit (0≤t≤a)

During pulse application C_e is approximately 200 $\mu\mu f$ (micromicrofarads).

After pulse, the charge of capacitor C equals the input pulse which reverse biases T_1. Assumed are an infinite emitter resistance and an emitter to base capacitance which is the result of an estimated 5-$\mu\mu f$ barrier capacity.[2] The following set of conditions is applied to the circuit, Fig. 2:

$I=15$ μa (microamperes) (constant)
$C_1=120$ $\mu\mu f$
$C_e=200$ $\mu\mu f$ for $0\leq t\leq a$
$C_e=5$ $\mu\mu f$ for $a\leq t\leq k^*$
$r_e=40$ ohms for $0\leq t\leq a$
$r_e=\infty$ for $a\leq t\leq k^*$
$r_b=400$ ohms
$\alpha=0.95$

where

r_b = base resistance of *2N168 A*
r_e = emitter resistance of *2N168 A*
C_e = base to emitter capacitance

Two applied input pulses, having an amplitude of 1 volt and pulse widths of 5 μsec (microseconds) and 0.25 μsec are used for this analysis.

Circuit Node Equations

For $0\leq t\leq a$, the node equations in the time domain can be written as:

$$\frac{e_{in}(t)-V_1}{r_b}+\alpha i_e+\frac{e_0(t)-V_1}{r_e}+C_e\frac{d[e_0(t)-V_1]}{dt}=0 \quad \text{at node no. 1}$$

$$\frac{V_1-e_0(t)}{r_e}+C_e\frac{d[V_1-e_0(t)]}{dt}-I-C_1\frac{de_0(t)}{dt}=0 \quad \text{at node no. 2}$$

$$i_e(t)=\frac{V_1-e_0(t)}{r_e}+C_e\frac{d[V_1-e_0(t)]}{dt}$$

Writing this equation in the Laplace transform domain, substituting for i_e, and eliminating V_1 between the two equations, $e_0(s)$ can be written as:

$$e_0(s)=\frac{e_{in}(s)[r_eC_es+1]-I(s)[r_e+r_b(1-\alpha)(r_eC_es+1)]}{r_er_bC_1C_e(1-\alpha)s^2+[r_e(C_1+C_e)+r_bC_1(1-\alpha)]s+1} \qquad (1)$$

Substituting $e_{in}(s)=1/s$, $I(s)=I_0/s$ into equation 1, and taking the inverse Laplace transform

$$e_0(t)=D+\frac{1}{2\sqrt{A^2-4B}}\times\left\{[D(A-\sqrt{A^2-4B})-2BE]\times e^{-\left(\frac{A+\sqrt{A^2-4B}}{2}\right)t}+[2BE-D(A-\sqrt{A^2-4B})]\times e^{-\left(\frac{A-\sqrt{A^2-4B}}{2}\right)t}\right\}$$

where

$$A=\frac{r_e(C_1+C_e)+r_bC_1(1-\alpha)}{r_er_bC_1C_e(1-\alpha)}$$

$$B=\frac{1}{r_er_bC_1C_e(1-\alpha)}$$

$$E=r_eC_e[1-I_0r_b(1-\alpha)]$$

$$D=1-I_0[r_e+r_b(1-\alpha)]$$

Substituting the numerical values for r_e, r_b, C_e, etc., the expression becomes

$$e_0(t)=1-0.53e^{-7.18\times10^{8}t}-0.47e^{-7.88\times10^{7}t} \qquad (2)$$

For $a\leq t\leq k$, the equivalent circuit is shown in Fig. 3. One may correctly assume that the charge of capacitor C_1 is the steady-state value of 1 volt because equation 2 reveals that the charge on the capacitor has essentially obtained its steady-state value in 0.05 μsec.

Writing the node equations for this circuit, substituting for e, and eliminating V_1 as was done in the previous circuit, the output voltage expressed in the transform domain is:

$$e_0(s)=\frac{C_ese_{in}(s)-I(s)[1+(1-\alpha)C_er_bs]+V_0C_1[1+(1-\alpha)r_bC_es]}{s[(1-\alpha)r_bC_eC_1s+C_e+C_1]} \qquad (3)$$

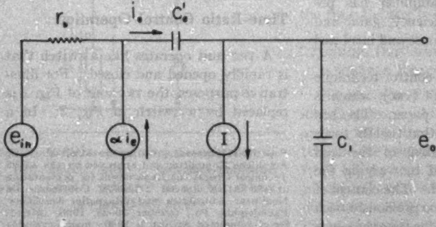


Fig. 3. Equivalent circuit (a≤t≤k)

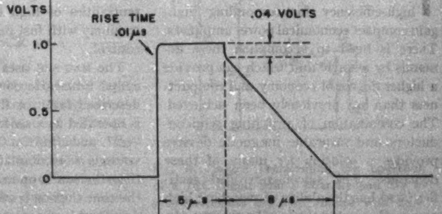


Fig. 4. Rundown waveform for a 1-volt 5-μsec pulse

Substituting $e_{in}(s)=0$, $I(s)=I_0/s$, $V_0=1$, and taking the inverse transform, $e_0(t)$ becomes

$$e_0(t)=\frac{B'I_0+D'C_1-A'D'I_0}{D'^2}-\left(\frac{I_0}{D'}\right)t+\left[\frac{D'^2-B'I_0-D'(C_1-A'I_0)}{D'^2}\right]e^{-\frac{D'}{B'}t} \quad (4)$$

where

$A'=(1-\alpha)r_bC_e$

$B'=(1-\alpha)r_bC_1C_e$

$D'=C_1+C_e$

$I_0=15\ \mu a$ (constant)

Using numerical values for r_b, C_e, C_1, equation 4 becomes

$$e_0(t)=0.96-1.2\times10^5t+0.04e^{-10^{10}t} \quad (5)$$

Experimental Results and Discussion

Two pulses of widths 5 and 0.25 μsec and an amplitude of 1 volt were applied to the rundown circuit, Fig. 1; since the waveforms obtained from experimental and theoretical plots of equations 2 and 5 agreed so closely it was possible to plot them as a single waveform, Fig. 4.

Also, in both cases the rundown waveform dropped approximately 0.04 volt after the trailing edge of the input pulse. This drop, resulting from C_1 discharging through the barrier capacitance and base resistance, decreases rundown time, lessening its sensitivity. The effect is minimized by an emitter-follower transistor with low barrier capacitance.

For gating purposes the rundown voltage is applied to a differentiating circuit obtaining a pulse width proportional to the rundown time. Fig. 5 illustrates the waveforms obtained after rundown and differentiating.

Conclusions

A 2-transistor rundown circuit, which converts low-level pulse heights to gates, produces a linear variation of voltage with time The linearity of rundown depends upon the constancy of the constant-current generator and ambient temperature. Because of inherent barrier capacity, the ideal waveform is unobtainable, but it may be closely approximated by using a transistor with low barrier capacity. For the circuit to perform accurately, the transistors must be kept at a constant temperature and the applied voltages must be well regulated. The circuit operates well from 0 to 3 volts; accuracies of 0.5% are possible using this circuit for pulse-height to pulse-width conversion together with a differentiating circuit. Achieving this accuracy means carefully selecting components and controlling temperature on all critical elements.

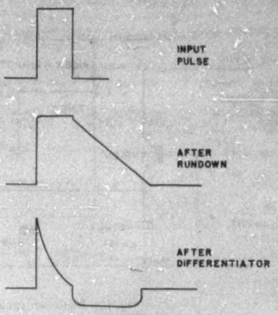


Fig. 5. Waveforms

References

1. Pulse and Digital Circuits (book), J. Millman, H. Taub. McGraw-Hill Book Company, Inc., New York, N. Y., 1956, pp. 579–82.

2. An Introduction to Junction Transistor Theory (book), R. D. Middlebrook. Chapman & Hall, Ltd., London, England, 1957, pp. 154–60.

A New Magnetic-Controlled Rectifier Power Amplifier with a Saturable Reactor Controlling On Time

RAY E. MORGAN
MEMBER AIEE

THERE IS AN increasing demand for high-efficiency fast-responding high-gain compact economical power amplifiers. There is need to accomplish these demands by a single unit which can provide a higher degree of economy and compactness than has previously been achieved. The combination of switching semiconductors and saturable magnetic devices provides a solution for many of these demands. The silicon-controlled rectifier (SCR) has the ability of switching high power (1 to 100 kw). Combining the SCR with saturable magnetic devices, a power amplifier (MAG-SCR) is provided. The MAG-SCR power amplifier has potentialities of high efficiency, gain, and economy with fast response and insulated control.

The MAG-SCR uses a control technique called time-ratio control (TRC), which is described later in this paper. The SCR is operated as a switch that rapidly turns "on" and "off." Control of the load voltage is accomplished by varying the time the SCR is on or off. The control of the time the SCR is off has previously been described.[1] Control of the time the SCR is on is desirable for many power-amplifier applications. The operating frequency remains approximately constant and the time of response to control signals is more constant, usually faster, than in applications where the time during which the SCR is off is controlled.

This paper presents a technique for controlling the MAG-SCR, Fig. 1, by varying the time the SCR is turned on with the use of an insulated control winding to provide the characteristic of Fig. 2. The principle of TRC and the technique of controlling "on time" by a saturable reactor are presented.

Time-Ratio Control Operation

A TRC unit operates like a switch that is rapidly opened and closed. For illustrative purposes, the TRC unit of Fig. 1 is replaced by a switch in Fig. 3. In a

---

Paper T-121, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers, Philadelphia, Pa., October 26–28, 1960. Manuscript submitted August 1, 1960; made available for printing February 21, 1961.

Ray E. Morgan is with the General Electric Company, Schenectady, N. Y.

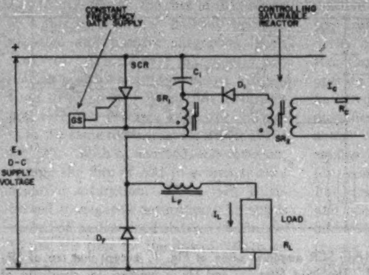


Fig. 1 (left). MAG-SC-R amplifier using TRC with constant frequency and variable "on time." The circuit is the same as described in reference 1, only saturable reactor SR_2 and diode D_1 are added

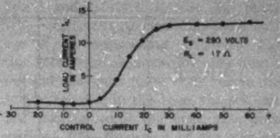


Fig. 2 (right). Transfer characteristic of MAG-SCR of Fig. 1

typical MAG-SCR, the switch S_a is opened and closed at approximately a constant rate, such as 3 kc. The length of time the switch is closed is varied, in order to control the load voltage.

Consider a switch controlling a resistive load as in Fig. 3. The average load voltage $E_L = E_S$ times the ratio of the time switches S_a is closed to the time S_a is opened and closed. For example, if S_a is closed 50% of the time, $E_L = 0.5\ E_S$, or if S_a is closed 10% of the time, $E_L = 0.1\ E_S$, etc. In Fig. 3 the ripple of the load current i_L is high and undesirable. Inductance L_F and diode rectifier D_F of Fig. 4 filter the load current providing a smooth direct current and smooth direct voltage E_L. Energy is stored in L_F when S_a is closed and L_F supplies the energy to the load when S_a is open.

The switch S_a of Fig. 4 is replaced by the SCR circuit of Fig. 1. Fig. 1 shows a circuit that uses the technique of controlling the length of time that SCR conducts by the control current I_C, which is described later. The transfer characteristic of the MAG-SCR of Fig. 1 is shown in Fig. 2.

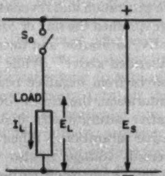


Fig. 3. Switch representing the operation of a TRC circuit with a resistance-unfiltered load. Load voltage $E_L = t_R E_S$ where $t_R = t_{ON}/(t_{ON} + t_{OFF})$

t_R = ratio of time
t_{ON} = time S is closed
t_{OFF} = time S is open

MAG-SCR Operation

The MAG-SCR operates like the switch circuits of Fig. 3 and Fig. 4 except that solid-state devices are used instead of a switch with contacts. The SCR of Fig. 1 is turned on by a pulse of voltage from the constant-frequency gate supply (GS). The SCR remains on (conductive) until the saturable reactor SR_1 reaches negative saturation. When SR_1 reaches negative saturation, capacitor C_1 discharges and reverses the voltage across SCR (explained later). Once the voltage of SCR is reversed (for approximately 20 microseconds), SCR remains turned off (nonconductive) until another pulse is received from the GS. The GS turns on the SCR at a constant rate, such as 3 kc, and the saturable reactor SR_1 turns off the SCR at various times after the SCR is turned on. Reactor SR_1 leaves the SCR on for a period of time t_{on} ranging from $t_{on} < 0.1\ t_{GS}$ to $t_{on} > 0.9\ t_{GS}$ when t_{GS} is the interval of time between pulses of the GS voltage.

The length of time t_{on} is controlled by the saturable reactor SR_2, Fig. 1. Reactor SR_2 operates similar to a conventional half-wave self-saturating magnetic amplifier.[2,3] Reactor SR_2 reacts upon reactor SR_1 and SR_2 controls the length of time required for SR_1 to reach negative saturation (discussed later in this paper). Control current I_C controls SR_2 and SR_2 reacts upon SR_1 resulting in I_C controlling the time the SCR is on.

"Turning Off" the SCR by a Saturable Reactor

The MAG-SCR of Fig. 1 operates by rapidly (approximately 3 kc) turning power on and off to the load circuit. The GS turns on the SCR and saturable reactor SR_1 turns off the SCR. The technique of the turning off of the SCR has previously been described.[1] This paper briefly reviews the turning-off technique to simplify the explanation of controlling the "on time" by saturable reactor SR_2 of Fig. 1.

The saturable reactor SR_1, capacitor C_1, and SCR of Fig. 1 are shown in Fig. 5. The tap of SR_1 is omitted for simplicity of explanation. The MAG-SCR is useful and practical with the tap of SR_1 omitted (discussed later). The waveshapes of the MAG-SCR of Fig. 5 are shown in Fig. 6.

The flux of reactor SR_1 is near negative saturation and $e_C = E_S$, with e_C positive at the dot, when SCR is turned on, due to previous operation of SR_1. After SCR is turned on, the flux of SR_1 rises toward positive saturation (e_1 is positive at dot of SR_1). While the flux of SR_1 is rising, $e_1 = E_S$ (supply voltage), capacitor C_1 voltage $e_C = E_S$ and the voltage drop of SCR is negligible, i.e., approximately 1 volt. After the flux of SR_1 reaches positive saturation, capacitor voltage e_C reverses, e_C changes from negative at the dot to positive at the dot, as explained later.

With SCR turned on, $e_1 = e_C$, neglecting the voltage drop of SCR. When SR_1 is saturated, a resonant circuit is formed by capacitor C_1 and the saturated inductance of SR_1. As C_1 discharges (e_C decreasing), the energy, 1/2 $C_1\ e_C^2$, of C_1 is transferred to SR₁ in the form of 1/2 $L_s\ i_1^2$ where L_s is the saturated inductance of SR_1 and i_1 has the polarity shown in Fig. 5. As i_1 decreases, the energy of SR_1 is returned to C_1 and e_C rises to almost $e_C = E_s$ except e_C is negative at the dot of C_1. When e_C reaches maximum

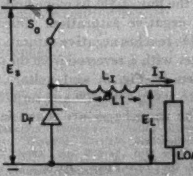


Fig. 4. Switch representing the operation of a TRC circuit with a filter L_F and D_F to provide smooth d-c load voltage with $E_L = t_R E_S$. When S_a is closed, L_F stores energy $= (E_S - E_L) I_L t_{ON}$ and when S_a is open, energy to the load is supplied by L_F of $E_L I_L t_{OFF}$

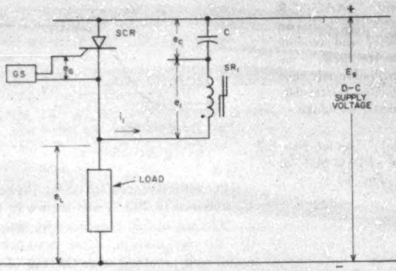


Fig. 5. Portion of MAG-SCR. Same as Fig. 1 except that SR_2 and D_1 are omitted. The tap of SR_1 is omitted for simplicity of explanation. The circuit is useful without the tap of SR_1, but remains at maximum-output voltage with SR_2 removed

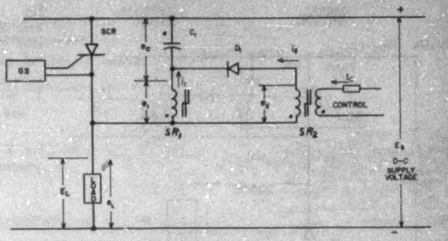


Fig. 7. MAG-SCR amplifier same as Fig. 1, except that tap of SR_1 is removed. The load filter L_F and D_F is omitted for simplicity of explanation but L_F and D_F are used in applications of MAG-SCR same as in Fig. 1. The load voltage E_l (average) is controlled by I_C similar to Fig. 2. Voltage e_L indicates instantaneous voltage as applied to the load filter of Fig. 1

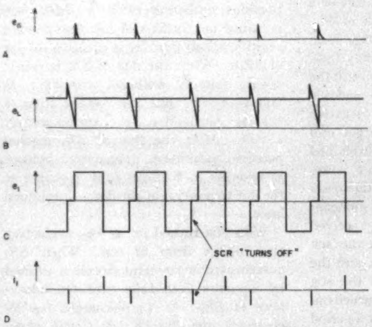


Fig. 6 (left). Waveshapes of the circuit of Fig. 5 same as complete MAG-SCR at maximum-output voltage

A—Gate supply voltage for SCR
B—Load voltage (without filter)
C—Voltage across SR_1
D—Current through SR_1 (i_1 is negligible when SR_1 is unsaturated)

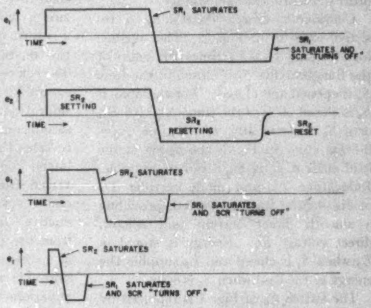


Fig. 8. Voltage waveshapes of Fig. 7

A—Voltage across SR_1 at load voltage E_L = maximum (same as Fig. 4)
B—Voltage across SR_2 at E_L = maximum showing SR_2 setting when e_1 is positive and SR_2 resets when e_1 is negative
C—Voltage across SR_1 at E_L = 1/2 of maximum
D—Voltage across SR_1 at E_L = 1/4 of maximum

voltage $e_C = E_s$, current i_1 attempts to reverse but as i_1 attempts to reverse, SR_1 becomes unsaturated. The flux of SR_1 moves toward negative saturation, i.e., e_1 is negative at dot of SR_1. With SR_1 unsaturated, i_1 is negligible (i_1 is the exciting current of SR_1) and e_C remains constant. Voltage e_C remains constant and i_1 negligible during the period of time for the flux of SR_1 to move from positive to negative saturation. Once the flux of SR_1 reaches negative saturation, C_1 discharges with i_1 reversed from the polarity shown in Fig. 5, and voltage e_C is applied to reverse the SCR and turn off the SCR. The waveshapes are shown in Fig. 6.

CONTROL OF "ON TIME" BY A SATURABLE REACTOR

The length of time the SCR conducts current, on time, is controlled by a saturable reactor. The MAG-SCR of Fig. 1 controls the load voltage E_L by the technique of TRC. The GS fires the SCR at a constant repetition rate. The load voltage is controlled by varying the on time of the SCR.

The on-time control is illustrated in Fig. 7, with voltage waveshapes shown in Fig. 8. The MAG-SCR of Fig. 7 is the same as that of Fig. 5, except that saturable reactor SR_2 is added. Reactor SR_2 is supplied by alternative voltage e_1 and SR_2 is reset each cycle of operation, such as 3 kc, to a flux level determined by control current I_C, Fig. 7. The amount of flux change from the flux level determined by I_C to the flux level at positive saturation determines the length of time the SCR is on.

To illustrate the operation of reactor SR_2, consider the MAG-SCR of Fig. 7 with $I_C = +50$ ma (milliamperes), the flux of SR_2 resets to negative saturation each time voltage e_1 is negative, see dot of Fig. 7. Reactor SR_2 has approximately the same volt-second rating as reactor SR_1. After the SCR is fired, the flux of SR_1 moves from the negative residual flux level to positive saturation and reverses capacitor voltage e_C; then the flux of SR_2 moves from negative saturation and turns off the SCR, as described above. While the flux of SR_1 moves from negative residual to positive saturation, the flux of SR_2 moves from negative saturation to slightly less than positive saturation. When voltage e_C is reversed, voltage e_1 also reverses (direction for D_1 to block current flow) and SR_2 is allowed to reset as shown in Fig. 8(A) and (B) where I_C supplies the reset. During this operation, $I_C = +50$ ma, reactor SR_2 has no effect upon the MAG-SCR. Reactor SR_2 determines the maximum on time which is designed for $E_L > 0.9\ E_s$.

When control current I_C is reduced to

$I_C = 15$ ma, Fig. 2, the flux of SR_2 resets to a flux level located halfway between negative and positive saturation levels (flux = 0). With $I_C = 15$ ma, reactor SR_2 reaches positive saturation before SR_1 has time to saturate. When SR_2 saturates positive, SR_2 reverses capacitor voltage e_C in the same manner as described above for SR_1. With e_C reversed, e_1 is reversed, forcing the flux of SR_1 to return toward negative saturation. When SR_1 is returned to negative saturation, the SCR is turned off as described above. With $I_C = 15$ ma, the time for the flux of SR_1 to return to negative saturation is half of the time of that required when $I_C = 50$ ma (voltage wave C compared with wave A of Fig. 8). The on time is reduced to half, and load voltage E_L is reduced accordingly. Likewise, as control current I_C is further reduced, the on time is reduced, Fig. 8(D). As I_C is varied, the on time varies and varies load voltage E_L Fig. 2.

The reaction of the tap of SR_1 in Fig. 1 changes the waveshape of voltage e_1 as illustrated in reference 1, but the operation of controlling the on time remains the same as described above. The reaction of varying the tap of SR_1 and other arrangements of SR_1 vary the characteristic of the MAG-SCR in reference to variations of load impedance, supply-voltage variations, etc. With any arrangement of SR_1, the reactor SR_2 controls the on time of the SCR and, in turn, controls the load of the MAG-SCR.

DESIGN FACTORS OF SATURABLE REACTORS

The designs of saturable reactors SR_1 and SR_2 of Fig. 7 are similar to those of saturable reactors of conventional magnetic amplifiers.[2,3] The inductances of SR_1 and SR_2, when saturated, are more important than the saturated inductance of conventional magnetic amplifiers. When SR_1 saturates, the saturated inductance of SR_1 and capacitor C_1 must provide the proper frequency that can apply reverse current and voltage to the SCR sufficiently long to turn off the SCR. The exact time to turn off the SCR depends upon the particular SCR, load current, supply voltage, ambient temperature, etc. The turn-off time for the SCR is usually between 10 and 40 microseconds.[1] The saturated inductance of SR_2 is usually similar to the saturated inductance of SR_1 but varies with different applications.

The other design parameters of SR_1 and SR_2 are similar to those of a conventional magnetic amplifier. Reactors SR_1 and SR_2 are designed to be able to make a flux excursion from negative to positive saturation and return to negative saturation during each complete cycle of operation of the MAG-SCR. Reactors SR_1 and SR_2 are designed to operate with the needed voltage, current, and frequency as described for magnetic amplifiers in references 2 and 3. Reactors SR_1 and SR_2 are usually much smaller and less expensive than reactors for conventional magnetic amplifiers. For example, a MAG-SCR controlling 5 kw and operating at 3 kc uses reactors SR_1 and SR_2 of about the same size as the reactor for a 5-watt magnetic amplifier operating at 60 cycles per second.

The exact design parameters of SR_1 and SR_2 vary with different application and a description of these parameters is beyond the scope of this paper. The design techniques described in references 2 and 3 are applicable to designing SR_1 and SR_2. The turn-off requirements of the SCR are shown in reference 1.

CONSTANT-FREQUENCY GATE SUPPLY FOR SCR

The GS to fire the SCR can be any oscillator circuit that supplies voltage pulses. The unijunction transistor oscillator previously described,[1] is one of the most frequently used GS circuits. The requirements for the GS circuit are:

1. A pulse of voltage and current must be provided for a sufficient period of time to turn on the SCR.
2. The length of time of the pulse must be less than the time t_{on} that the SCR is on.
3. The GS frequency must be relatively constant.

The requirement of voltage, current, and time to fire the SCR depends upon the SCR that is being used and the range of temperature of the SCR junction. Specifications for firing requirements of General Electric SCR's are shown in reference 1. For example, the SCR type *C-35* can be fired by 3 volts 80 ma applied for 5 microseconds at -60 C (degrees centigrade). At industrial ambient temperatures less current is required.

The length of time T_{GS} that the GS pulse must be applied to the SCR is not critical. Time t_{GS} must exceed the time required to fire the SCR. but $t_{GS} < t_{on}$ [min (minimum) time SCR is on]. If $t_{GS} > t_{on}$, the SCR turns on immediately after turning off and the output load voltage becomes maximum instead of remaining reduced.

The frequency of the GS must be relatively constant, usually ±5%. The time of a cycle of the GS oscillator, the time t_{PP} between pulses, must remain with $t_{PP} > t_{on}$ [max (maximum) time SCR is "on"] and $t_{PP} < 1.1\ t_{on}$ (at t_{on} max). If $t_{PP} < t_{on}$ (max), as maximum load voltage E_L is approached ($E_L = 0.9\ E_S$), the gate pulse occurs while the SCR is on and the gate pulse ends before the SCR turns off. In this case, the SCR cannot turn on until the next gate pulse arrives. The frequency of the MAG-SCR drops to half the frequency of the GS and the load voltage E_L drops to $E_L = 0.45\ E_S$ (supply voltage) instead of $E_L = 0.9\ E_S$. This provides an undesirable characteristic. If time $t_{PP} > 1.1\ t_{on}$, the maximum load voltage $E_L < 0.9\ E_S$ causing excessive reduction in the maximum-load voltage. The time t_{PP} should be held between $t_{PP} > t_{on}$ (max) and $t_{PP} < 1.1\ t_{on}$ (max) for best results. The tolerance on t_{PP} may be relaxed at the expense of reduced maximum load voltage. There are techniques that remove the requirement for holding constant frequency, t_{PP} constant, without sacrifice of maximum load voltage. The unijunction transistor gate supply[1] is satisfactory for the MAG-SCR to provide load voltage $E_L < 0.1\ E_S$ (supply voltage) to $E_L < 0.9\ E_S$.

Conclusions

The MAG-SCR, with a saturable reactor to control "on time," provides a compact, fast-responding, efficient, and economical power amplifier. Presently available controlled rectifiers with practical, economical saturable reactors provide an excellent power amplifier for loads of 0.5–50 kw. Greater load power is controlled with multiple seriesing and paralleling of SCR's. The MAG-SCR has advantages at low-load power, 100 watts and less, where the supply voltage exceeds 100 volts or the ambient temperature exceeds 100 C. The MAG-SCR with control of on time provides a faster response than the MAG-SCR with control of off time, reference 1. On-time control operates at a constant frequency in contrast to off-time control that requires large reduction in operation frequency to obtain reduced output voltage.

References

1. CONTROLLED RECTIFIER MANUAL. F. W. Gutzwiller, D. V. Jones, H. R. Lowry, T. P. Sylvan, E. E. Von Zastro. General Electric Company; Liverpool, N. Y. "Morgan Circuit," pp. 119–26, "Unijunction Transistor Firing Circuits," pp. 50–54; "General Electric Controlled Rectifier Specifications," pp. 204–43; first edition, 1960.

2. MAGNETIC-AMPLIFIER CIRCUITS, (book), W. A. Geyger. McGraw-Hill Book Company, Inc., New York, N. Y. "Basic Single-Core Internal-Feedback of Magnetic Amplifiers," pp. 146–59; "Basic Two-Core Internal-Feedback of Magnetic Amplifiers," pp. 160–71; first edition, 1954.

3. MAGNETIC AMPLIFIERS (book), H. F. Storm. John Wiley & Sons, Inc., New York, N. Y. "Half-Wave Amplistat," pp. 309–19; "A-C Amplistat," pp. 232–53; first edition, 1955.

Pulse Discharges Preceding Sparkover at Low Voltage Gradients

G. W. PENNEY
FELLOW AIEE

S. CRAIG
NONMEMBER AIEE

BREAKDOWN of gases between clean metallic electrodes has been studied extensively and has resulted in an acceptable explanation for sparkover where the gradient is of the order of 30 kv/cm (kv per centimeter) or more, but it does not explain the sparkovers which occasionally occur in high-voltage apparatus where the gradient appears to be nearer 5 than 30 kv/cm. A previous paper[1] has shown that for negative corona, sparkover voltage is a function of the resistivity of any coating on the anode or passive electrode and can be especially low if the resistivity is high. This type of sparkover was found to be preceded by repetitive current pulses. Each pulse was found to have the same current-time characteristics and the same velocity of propagation as the single prebreakdown streamer described by Meek and Craggs[2] and Parks and Cones,[3] but which were produced with clean electrodes and at higher gradients.

The work with high-resistivity anode coatings indicated that there was a mechanism by which one pulse can cause another, thus resulting in build-up of successive pulses and sparkover in regions of much lower gradient. The tests suggested that this type of progressively increasing pulses and resulting sparkover could occur wherever there was an imperfect insulating film, a field of 5 kv/cm or more, and some source of negative ions. If this is true, these phenomena can account for the occasional sparkover in high-voltage apparatus at relatively low voltage gradients. Because of the importance of high-voltage sparkover in electrostatic precipitators and all other types of high-voltage apparatus, experiments were devised to determine the exact conditions which must be present to produce the presparkover pulsed discharges.

These discharges, which consist of repetitive current pulses and which we have called "flares," were produced in three different types of experiments, each of which was used to study a particular aspect of the over-all phenomena:

1. With positive point-to-plane corona small repetitive pulses of voltage can cause large pulses of current and sparkover at relatively low voltages. This type of experiment was useful for studying the effect of pulse frequency, the velocity with which the discharge propagates in the field, and the conditions which lead to sparkover. However, there had to be corona prior to the pulse. The resulting space charge prevented an accurate evaluation of the exact field configuration which could cause the current pulses.
2. In electric apparatus a hole or defect in an insulating film appears to be a likely cause of flares and sparkover provided there is a source of negative ions in the region. Measurements of the potential of the surface of an insulating electrode covering were made which showed the presence of a radial voltage gradient on the surface around a hole in the covering. However, these experiments did not permit the desired control of the radial gradient and hence did not show directly its effect in starting the flare current pulses.
3. The final experiment used parallel plane electrodes in which the central area of the anode consisted of a system of concentric rings insulated from each other. The potential of these rings could be controlled to produce a radial gradient and a minute repetitive spark at the center. This did produce flares and sparkover and permitted a more detailed study of the fields needed to produce this type of discharge.

Flare

The term flare is used to denote a discharge consisting of repetitive current pulses or streamers which originate from a fixed location on the anode and frequently tend to increase in magnitude until sparkover occurs. Each current pulse appears to be due to ionization in a very much localized high-field region which propagates from anode to cathode. The current pulses tend to repeat naturally about every 33 μsec (microseconds), although there is some random variation in the exact period between consecutive pulses. Each pulse rises to a peak in 0.05 μsec and decays again to a low value in about 0.5 μsec. An oscillogram of the current pulse is shown in Fig. 1. The average d-c value of the flare discharge current is typically several hundred μamp (microamperes), and the peak value of the current pulse is typically 30 to 80 milliamperes. The integrated current pulse accounts for only about one third of the total average current.

Since ionization by collision is accompanied by visible light, the light emitted by the discharge was studied and found to consist largely of light pulses moving from anode to cathode with a velocity of about 10^7 cm/sec. This agrees with previous measurements of the velocity of streamers in positive point-to-plane corona.[4]

The light from a single pulse is so faint that as yet no satisfactory photograph has been obtained showing only one pulse. Photographs of a few pulses show only fine lines with all paths common near the anode. In most cases paths are common for the first 1/4 inch but beyond this point the paths tend to diverge, so that in a long-exposure photograph, or as seen by the eye, the discharge appears to flare out in a conical form. A 2-sec time exposure in which the paths of approximately 60,000 pulses are superposed is shown in Fig. 2.

The first flares studied were produced by using corona with a wire as the active electrode and a large cylinder as the passive electrode. These flares were

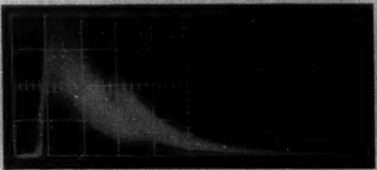


Fig. 1. Oscillogram of flare current pulse. Exposure 3 sec with about 30,000 pulses/sec. Vertical, 20 ma/square; horizontal, 0.1 μsec/square

---

Paper 61-91, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 20, 1960; made available for printing December 8, 1960.

G. W. Penney is with the Carnegie Institute of Technology, Pittsburgh, Pa., and S. Craig is with the General Electric Company, Syracuse, N. Y.

This work was supported in part by National Institutes of Health grant *RG 6402*, and in part by National Science Foundation grant *G2424*, and was done in part by Samuel B. Craig in partial fulfillment of the requirements for a Ph.D. degree at the Carnegie Institute of Technology.

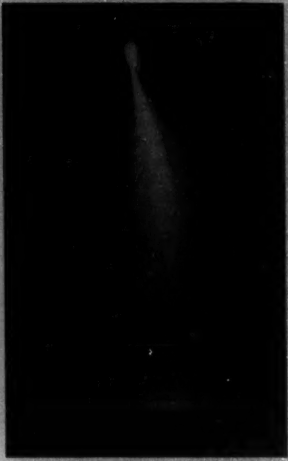


Fig. 2. Flare discharge. Exposure 2 sec with about 30,000 pulses/sec

produced by insulating material on the anode (on the wire with a positive corona or on the passive electrode with negative corona). Probably the flares of widest interest are those which may occur between large electrodes where there is no apparent source of corona. Flares and sparkover have been produced between large electrodes with an insulating film on the anode and various sources of ionization to trigger an initial spark. This work has shown that flares can be produced in this way but the detailed conditions which result in the succession of pulses are only partly determined. Hence no detailed report of this work can be given at this time.

This paper describes two types of experiments in which flares and sparkover have been produced with the use of clean metallic electrodes. To produce flares from clean metallic electrodes, successive voltage pulses must be superimposed on the applied d-c potential. These experiments have given information on the fields required to produce a flare and have made it possible to determine the effect of varying pulse frequency. However, the experiments did not determine how the necessary frequency is produced when the flare occurs with insulating material on an anode.

Flares With Positive Points and Pulsed Voltages

While the flare discharge was first observed in corona with insulating material on a positive electrode, it was found early in the course of the investigation that the flare would also result with positive corona at a positive point electrode without insulating material if some pulsation at a rate near 30 kc was added to the d-c high voltage applied to the point. In particular the flare will occur at a positive point if a very small gap is connected in series with the point. Breakdown of this series gap then results in a slight but sharp increase in the total voltage applied to the point. The flare will occur only if the series gap is short enough to break down repeatedly at a rate near 30 kc.

To determine the effects of some of the separate features of voltage pulsation that are important in producing the flare, a pulse-generating circuit was used with a positive point-to-plane system. An oscillogram of the voltage pulse produced by the pulse-generating circuit is shown in Fig. 3. This voltage pulse, or exciting pulse, was used to produce the flare discharge by adding it to the d-c high voltage applied between the point and plane.

Fig. 3. Oscillogram of voltage pulse produced by hydrogen thyratron circuit. Vertical, 200 volts/square; horizontal, 0.1 μsec/square

Preliminary work showed that a high rate of rise of voltage was essential but that the required value of the pulse was only 400 to 800 volts as compared with a typical d-c component of 25 kv. The polarity of the pulse must be such that the total voltage is increased by the fast rise of the pulse. Fig. 4 shows the pulse generator circuit and its connection to the electrode system. If sufficient direct voltage is applied to the point, current pulses similar to that of Fig. 1 can result from the exciting voltage pulses.

The possibility of forming a current pulse depends on the interval between the successive exciting voltage pulses. Below 30 kc a current pulse results from every exciting pulse. However, if the exciting pulse frequency is raised to 50 kc, it is found that a current pulse results from some of the exciting pulses but not from all of them. The current pulses of

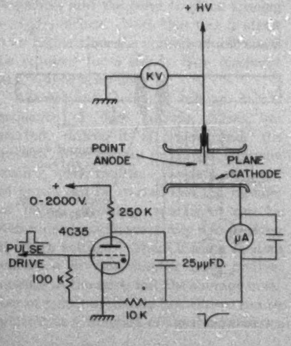


Fig. 4 (left). Hydrogen thyratron pulse circuit connected to point-to-plane electrode system

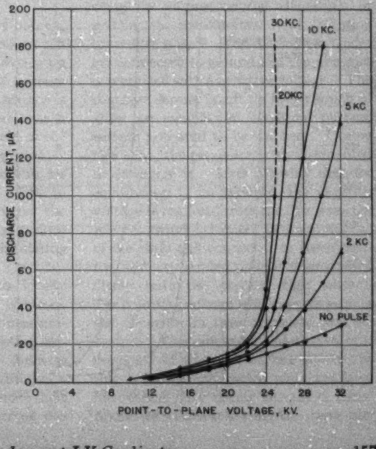


Fig. 5 (right). Current versus voltage for positive point-to-plane electrode system, showing effect of exciting pulse repetition rate

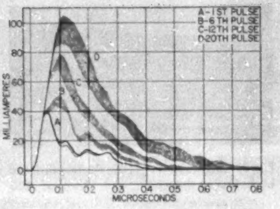


Fig. 6. Reproduction of oscillograms showing build-up of current pulses during a burst of 25 pulses. Shaded areas show range of values measured during a number of trials

the flare cannot be made to repeat regularly at an average rate exceeding about 40 kc. If the exciting pulse repeats at 60 kc, a current pulse will usually result only from every other exciting pulse.

If average discharge current is measured as a function of direct voltage and exciting pulse frequency, the current is relatively stable over a considerable range of voltage and frequency, but critical regions are found where the current tends to rise until a spark occurs. A typical family of these curves is shown in Fig. 5. Analysis of the portion of these curves represented by the solid lines and experimental points shows that at any given voltage the current increases as a linear function of frequency. This would indicate that at a given voltage all current pulses have the same magnitude, so that as frequency is increased a linear increase in current results.

The dashed portion of the curve for 30 kc in Fig. 5 shows the characteristic of a large increase in current for a relatively small change in voltage. It is only over this steeply rising portion of the curve that the glow of the flare is exhibited. The unstable nature of the flare results in sparkover if a constant direct voltage is maintained, but the flare can be observed as a steady discharge with only infrequent sparking by using a high-impedance voltage source. The direct current and voltage then vary somewhat at random so that exact points cannot be plotted. The sparking which accompanies the flare becomes more frequent as the average direct current is increased.

At the lower pulse repetition rates the characteristic of a nearly constant voltage over a wide range of currents is not found, the intensity of the glow is greatly reduced, and a considerably higher voltage is required for sparkover than at the 30-kc rate.

If similar families of current-voltage curves are plotted for other point-to-plane spacings, it is found that the voltage for a given current on the steeply rising part of the curve is very nearly a linear function of the point-to-plane spacing, within the range of 1/2 inch to 3 inches. This voltage-spacing relation shows almost no dependence on the exact electrode configuration in which the flare is produced. Hence the flare and accompanying sparkover seem to be associated with an average voltage gradient of about 5 kv/cm for gaps of a few inches in length. A departure from this figure might be found at larger spacings, which were not investigated in the present work.

Build-up of Pulses in Brief Prebreakdown Flares

The sparking which accompanies the flare discharge is evidence of prebreakdown currents increasing to a higher value, suggesting a build-up in the amplitude of successive current pulses and consequently a dependence of a pulse on the one preceding it. The occurrence of such an amplitude build-up is confirmed by examination of the current caused by brief prebreakdown flares which precede sparkover in negative corona with high-resistivity material on the passive electrode. Since the duration of such prebreakdown flares is short, the current actually takes the form of a sequence or "burst" of flare current pulses, with the sequence being terminated by a spark.

The individual current pulses in such a sequence are not all of the same amplitude; the first pulses of the sequence are small and a gradual build-up in amplitude takes place. A spark appears to result when the current pulse amplitude reaches a peak value of about 100 milliamperes. The first pulse of a sequence may have a peak amplitude of only a few milliamperes.

To study this amplitude build-up under controlled conditions, brief flares were produced in the positive point-to-plane configuration by applying the exciting voltage pulses in sequences of a given number rather than continuously. In a specific case each sequence consisted of 25 exciting pulses, with the exciting pulse frequency remaining at 30 kc during each sequence. The bursts or sequences are repeated at a rate of only about 50/sec, so that no sequence was affected by the preceding one. The display of an in-

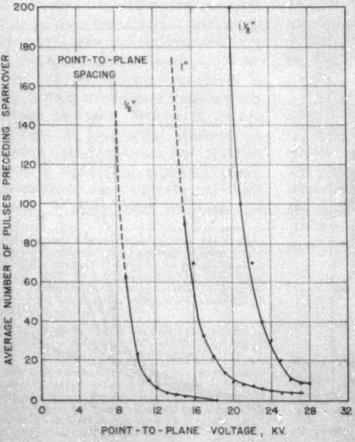


Fig. 7 (left). Average number of pulses preceding sparkover versus voltage, for three point-to-plane spacings

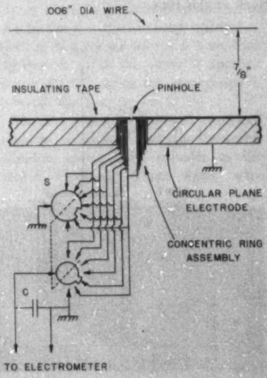


Fig. 8 (right). Circuit and electrode arrangement for measuring surface potential of insulating layer in presence of corona

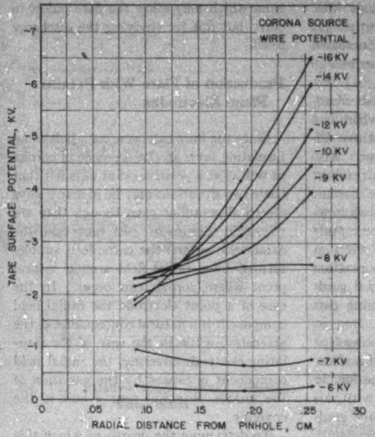
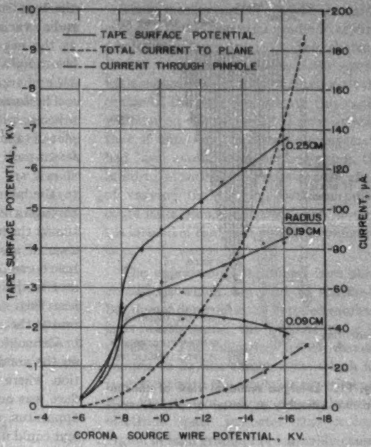


Fig. 9 (left). Potential of insulating tape surface versus radial distance from pinhole for apparatus of Fig. 8

Fig. 10 (right). Surface potentials and currents for apparatus of Fig. 8



dividual pulse at any place in the sequence is made possible by using the delaying sweep feature of the Tektronix model *535* oscilloscope.

Fig. 6 is based on oscillograms of the 1st, 6th, 12th, and 20th current pulses of each sequence. The build-up in amplitude is clearly evident. Each oscillogram was a time exposure photograph of the oscilloscope screen and included about a thousand current pulses, each of which occurred in a separate sequence. For each picture the pulse shown always appeared at the same place in the sequence. In each oscillogram the amplitude of a particular current pulse repeated fairly closely from one sequence to the next. The ringing or oscillation shown on the oscillograms was caused by the exciting voltage pulse and was not a part of the discharge current. The d-c high voltage was held constant at 24 kv throughout the sequences and the point-to-plane spacing was $1^1/_2$ inches. Note that this is about 5 kv higher than the voltage which would be required for a steady flare discharge at $1^1/_2$-inch spacing.

Above some particular voltage, once a sequence of flare current pulses has started, unless it is interrupted, the growth of the current pulses will continue until a spark results. Hence a sequence of pulses can be started and the number of pulses preceding sparkover can be determined oscilloscopically for a given d-c high voltage. There is some random variation in this number from one sequence to the next, but the average number of pulses preceding breakdown can be plotted as a function of the d-c anode-to-cathode voltage, as is done in Fig. 7 for three point-to-plane spacings.

Referring to the curve for $1^1/_2$-inch spacing, the number of pulses required before breakdown appears to increase indefinitely as the voltage is reduced toward a value of about 19 kv. For voltages below this value no build-up in pulse amplitude takes place and sparkover does not result. As the voltage is increased from this value, the amplitude build-up occurs at an increasing rate so that a smaller number of pulses is required preceding a spark.

Surface Potential of Insulating Electrode Covering

When flares occur in negative corona, they originate at some discontinuity such as a hole in a high-resistivity covering on the passive electrode. Since the average field over the passive electrode surface is generally much lower than that near a point electrode, a locally high field at the discontinuity in the insulating surface layer would appear to be responsible for the starting of the flare pulses. Due to the corona, charge can collect on the insulating surface and result in surface gradients near any hole in the insulating layer.

To determine the conditions locally near such a discontinuity, an electrode assembly was made to allow the measurement of the potential of an insulating surface in the presence of corona. Because no angular variation in the potential of a plane surface is expected around a hole in the insulating layer, the electrode was made in the form of a circular plane with an assembly of insulated concentric rings at its center. A piece of Scotch no. *33* electrical tape was placed on the plane electrode surface with a pinhole in the tape at the center of the concentric ring assembly. A fine wire was placed parallel to the plane above the tape to serve as a source of negative ions. The arrangement of the electrodes and the circuit used to measure the charge on the tape are shown in Fig. 8.

The method of measurement essentially consists of transferring the charges on the tape surface above one of the rings to a large capacitor by applying a grounded plate to the tape surface, and then measuring the voltage on the capacitor. In making the measurements the following procedure is used: First all ring electrodes are connected to ground and the fine wire is energized with a negative voltage. The negative corona is left on long enough to allow an equilibrium condition of tape surface potential to be reached. During this time, the capacitor is short-circuited to discharge it. After allowing time for equilibrium to be attained, the negative voltage is removed from the fine wire, the selector switch is rotated to unground one of the rings and connect it to the capacitor, and a grounded plate is then pressed firmly onto the exposed tape surface. The resulting voltage which appears across the capacitor is then measured and recorded. This allows calculation of the potential of the tape surface above the ring before the grounded plate has been introduced. By taking successive measurements with each ring, the tape sur-

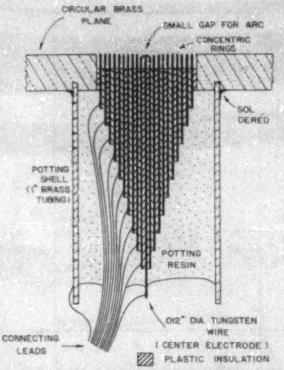


Fig. 11. Enlarged sectional view of concentric ring assembly at center of planar anode

face potential can be found as a function of radial distance from the pinhole.

Curves showing tape surface potential as a function of radial distance from the pinhole are shown in Fig. 9 for different values of fine wire voltage, and hence for different values of negative corona current. For small values of ion current to the tape surface there is relatively little radial variation in surface potential, but as the negative ion current is increased the radial surface gradient may exceed 20 kv/cm.

The same information presented in another way is shown in Fig. 10, which is a plot of tape surface potential at three different radii versus fine wire voltage. Also shown are curves for the total current to the plane electrode and the current through the pinhole only, both as a function of the fine wire voltage. The appearance of a significant current at the pinhole corresponds to the onset of ionization at that point. The radial surface gradient then increases with increasing current at the pinhole.

Although the punctured insulating tape on the anode is representative of a situation where flares can occur, the steady flare was not actually observed with this apparatus, since the radial surface gradient could not be controlled independently of the anode-to-cathode voltage. When local conditions at the anode were such that flare pulses could start, the anode-to-cathode voltage was high enough to cause a rapid build-up in the pulse amplitude, resulting in sparkover instead of the steady flare discharge. The apparatus was designed for determining some of the conditions which may allow the flare pulses to start, but not for studying the discharge itself.

Production of Flare With Parallel Plane Electrodes

Since the flare can occur at a hole in an insulating layer in the presence of corona as well as at a positive point, it is felt that these two situations may be in many ways electrically similar even though they are physically dissimilar. In each case the axial field toward the cathode is supplemented by a radial field around a local point where corona can occur. In the case of a point electrode the radial field component is a natural consequence of the geometry, while in the case of the insulating electrode covering, the radial field component is caused by the variation of the surface potential near a hole in the covering.

To determine the degree to which this radial field component is necessary for the propagation of flare pulses and to allow the radial and axial field components to be controlled independently, an electrode system was used consisting of two circular parallel planes with an assembly of insulated concentric rings at the center of the anode. This concentric ring assembly

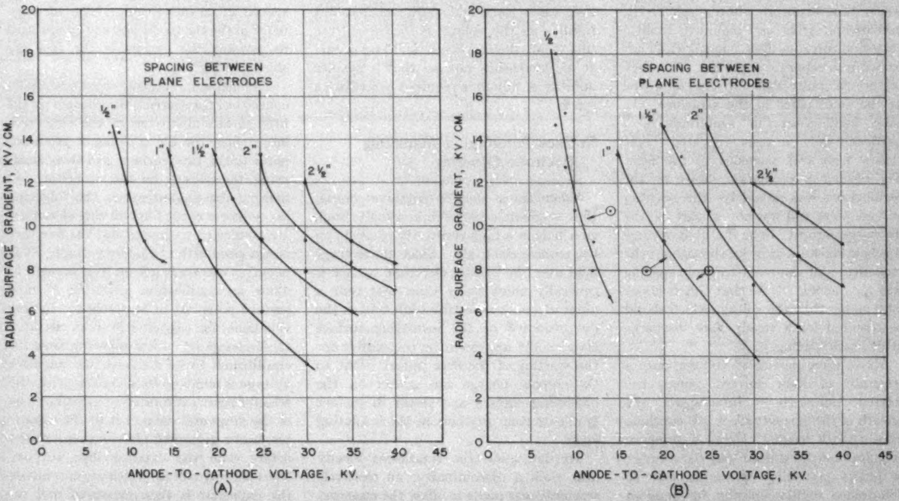


Fig. 12. Necessary radial gradient for pulse propagation versus anode-to-cathode voltage. Radial gradient shown is average between mean radii of two smallest rings

A—Gradient extends over all ten rings
B—Gradient extends only to sixth ring

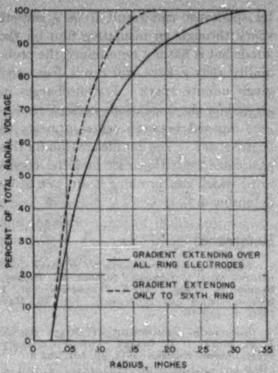


Fig. 13. Relative anode surface potential versus radius

consisted of ten thin brass rings with adjacent rings differing in radius by 1/32 inch. By applying successively different voltages to the rings, a radial surface gradient which is known and adjustable can be produced. Hence, no insulating covering is necessary and only metal electrodes are involved. Fig. 11 shows the details of the electrode ring assembly which forms the center of the anode of the parallel plane gap. The center electrode of the ring assembly is a wire inside the smallest ring. This ring is shaped so that application of the voltage pulse between it and the center wire at the 30-kc rate causes minute repetitive sparking at the center of the ring assembly and in the plane of the anode surface.

This parallel plane assembly showed that three conditions were necessary for the propagation of flare pulses to begin: (1) a minute spark, repetitive at 30 kc, at the center of the anode ring assembly; (2) sufficient radial gradient of the surface potential around the minute spark; and (3) sufficient axial field due to the anode-to-cathode potential difference. As the anode-to-cathode voltage is raised, less radial gradient is required to start the successive flare pulses which result in sparkover. The relation between the anode-to-cathode voltage and the required radial gradient for sparkover is shown by Fig. 12. The anode-to-cathode voltage is measured from the center of the ring assembly, which is the most positive point in the system. In Fig. 12(A) the radial gradient extended over all ten rings while in Fig. 12(B) it extended only to the sixth ring. In each case the value of the radial gradient given by the curve is the average between the two rings nearest the center. The forms of the potential variation over the rings are shown in Fig. 13.

Comparison of the curves of Figs. 12(A) and (B) shows that the required radial gradient at the center of the anode is not radically different whether the gradient extends to the tenth ring or only to the sixth ring. Hence it appears that a high radial gradient extending only 1 or 2 millimeters from the center of the anode might be sufficient to allow the flare pulses to start and result in sparkover. While only two forms of radial variation in anode surface potential were tried in the present work, it is expected that many other forms may be equally effective, since the form of potential variation does not appear to be critical provided that the gradient surrounding the minute spark is high enough.

Discussion of Experiments

Previous experiments with high-resistivity material on the anode in negative corona showed that for sufficient corona current a pulsating discharge called a flare can start at a location on the anode and result in a spark from anode to cathode. Usually this pulsed prebreakdown discharge is too brief to permit detailed observation, but occasionally it can be maintained long enough to record the variation of the current with time. The current consists of pulses of the form shown by the oscillogram of Fig. 1, with an average interval between pulses of about 33 μsec. These prebreakdown pulses can originate at a discontinuity in an insulating layer on the anode surface and, if the average voltage gradient is at least 5 kv/cm, can successively increase in amplitude until a spark results. Sparkover preceded by these repetitive current pulses has also been produced between relatively parallel electrodes with an imperfect insulating layer on the anode and a source of negative ions.

In the present work, particular aspects of the flare discharge were studied in separate but related experiments in order to determine conditions under which repetitive prebreakdown pulses will start and result in sparkover. The build-up of successive pulses was investigated: a positive point-to-plane electrode configuration was used with triggering voltage pulses superimposed on the applied d-c potential. Information regarding the conditions near a hole in an insulating layer on an electrode surface in the presence of corona was obtained by measuring the potential of the exposed surface of the insulating layer.

These measurements showed the existence of a radial potential gradient around a hole in the layer and led to the construction of a parallel plane electrode system with an assembly of concentric rings at the center of the anode to imitate this radial gradient on the surface. It now appears that if this final electrode arrangement had been available at the beginning, the build-up of successive pulses and the effect of varying the interval between pulses could have been studied with this apparatus instead of the point-to-plane configuration.

The experiments described here show conditions which can lead to sparkover but do not determine the detailed mechanisms which cause the current pulse to develop. The following hypothesis attempts to outline mechanisms which are consistent with the available experiments. The spot of light moving across the interelectrode space indicates that a local region of high gradient presumably caused by space charge distortion causes local ionization by collision. The velocity of propagation can be accounted for by assuming that radiation is a primary mechanism. These are the usual assumptions used to explain streamers. The decay of current to a low value in 0.5 μsec seems best explained by assuming that the ionization occurs primarily in a very narrow channel so that the ion density is high and the initial recombination very rapid. Electron attachment also tends to give a rapid reduction in current.

If photons of sufficient energy are produced in the region where ionization by collision is occurring, ions will be produced in all directions but those produced in the direction of the cathode will be in a region where the main field adds to the local field. This can be a mechanism causing the pulse to proceed in the direction of the cathode. The velocity of propagation of the pulse toward the cathode is about the same as the velocity with which electrons in the high field spot would be moving toward the anode. This in turn suggests that the velocity of propagation is determined by the rate at which electrons can be removed from the spot in which the ions are formed.

All of these mechanisms appear to occur both in streamers formed in regions of higher field and in the flares which may occur in gradients as low as 5 kv/cm. The distinctive feature which seems to be essential for causing a current pulse at the lower average field is a high radial field in the region immediately surrounding the point where the local minute spark or burst of ionization occurs. In a region surrounding the initial spark, free electrons will be produced by photoionization. A radial field draws more of these free electrons into the high field region where ionization by collision can occur.

Conclusions

Experiments indicate that with an imperfect insulating coating on the anode, and a source of ionization, current pulses followed by sparkover can occur if the average voltage gradient is of the order of 5 kc/cm. The individual current pulses observed here exhibit the same rate of rise, rate of decay, and velocity of propagation as the streamers observed by others in regions of higher gradient in which a single streamer tended to be followed by a sparkover. The work described here was intended to determine conditions which could cause this type of current pulse to start under conditions of relatively low gradient, and the conditions which can result in successive pulses building up until sparkover occurs. This succession of pulses starting from a single spot on the anode has been called a flare because of its characteristic appearance.

Sparkover for electrode spacings of 1 to 3 inches seems to occur at an average gradient of about 5 kv/cm for three quite different conditions: (1) Pulsed positive point-to-plane electrode configuration where the discharge originates in the high field region. (2) Parallel electrodes where the gradient is initially quite uniform except in the small region where the pulse starts. (3) Negative corona where the discharge starts at the anode or passive electrode which for low corona currents is a region of low field strength. The fact that sparkover results at about the same average gradient for these three conditions suggests that space charges resulting from the flare produce a gradient which is favorable to the development of a spark, or at least alter the field, so that the initial voltage distribution is rather unimportant.

This work shows that the flare followed by sparkover can result from a minute spark through an insulating film on the anode but it has not determined the probability that a flare will result from any given minute spark. For discharges in air and for the dimensions studied, the 33-μsec interval between pulses appears to be necessary. The detailed reason for this particular repetition rate and the conditions which can produce it are yet to be determined.

References

1. Sparkover as Influenced by Surface Conditions in D-C Corona, **Gaylord W. Penney, Samuel E. Craig.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, May 1960, pp. 112–18.

2. Electrical Breakdown of Gases (book), **J. M. Meek, J. D. Craggs.** Oxford University Press, London, England, 1953, pp. 201, 252 ff.

3. Surge Voltage Breakdown of Air in a Nonuniform Field, **J. H. Parks, H. N. Cones.** *Journal of Research*, National Bureau of Standards, Washington, D. C., vol. 56, no. 4, Apr. 1956, p. 201.

4. Fast-Time Analysis of Intermittent Point-to-Plane Corona in Air, **M. R. Amin.** *Journal of Applied Physics*, New York, N. Y., vol. 25, 1954, pp. 210, 358.

Commutation and Destructive Oscillation in Diode Circuits

ISTVAN SOMOS
MEMBER AIEE

SINCE the introduction of single crystal rectifier cells, the switching action from the forward- to the reverse-biased condition, in other words commutation, frequently has been analyzed. In the early stages of germanium and silicon rectifiers, destruction of cells was often caused by this commutation action. After the application of parallel capacitors or filtering circuits, the troubles were eliminated, and correctly designed filtering capacitors provided a simple, cheap, and effective way to assure trouble-free operation.

Further investigation showed that the need for protective devices depends also on the circuit parameters, and below certain values no protection was required. Furthermore, it was found that there are rectifier cells which in no practical cases require protection, because their behavior during commutation is similar to that of the selenium rectifier cell.

The investigation and testing of these phenomena and the study of how commutation may cause destruction of unprotected cells is the subject of this report. The tests were made on 9/16-inch-diameter alloyed p-type silicon rectifiers of 1.6-cm^2 (centimeters squared) junction area. The cells were a product of the General Electric Company, Low Voltage Switchgear Department, in Collingdale, Pa.

Commutation in Silicon Rectifiers

Fig. 1 shows a simplified illustration of commutation obtained in single-phase half-wave a-c d-c circuits with ohmic load. During the commutation, the rectifier cell is subjected to a cleaning-out process during which the injected carriers are removed from the semiconductor. During this period, the cell is in neither the forward nor the reverse condition.

In the first part of the commutation process, between points C and K in Fig. 1, the cell acts as a source of energy having a small internal resistance. This develops an electromotive force (emf), which has the polarity, and in the first instance the magnitude, of the forward voltage occurring across the junction. During this time, the entire circuit composed of the transformer, cell, and in certain cases the load, is equivalent to a common a-c circuit without rectifying action. The transformer voltage in the reverse period and the emf developed by the cell have the same polarity. In the forward period, the two voltages are directed against each other.

When the circuit contains inductance, an emf will be induced by it also; therefore the total emf of the circuit equals the sum of three partial emf. They are: the emf of the transformer, $V_{tr} = V_{tr}$ peak sin ωt; the emf V_j, developed by the rectifier; and the emf induced by the in-

---

Paper 61-89, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 3, 1960; made available for printing December 5, 1960.

Istvan Somos is with the General Electric Company, Collingdale, Pa.

The author gratefully acknowledges the technical assistance given by P. J. Colleran, W. H. Tobin, A. L. DeCecco, and W. Warburton; the fundamental work in commutation and circuitry by G. R. Lezan; the discussions in the General Electric Research Laboratory with E. M. Pell, R. N. Hall, J. J. Tiemann, H. M. Tweed in the matter of solid state physics; and the discussions in the General Electric Low Voltage Switchgear Department with C. E. Rettig, G. J. Bower, R. P. DePuy in the matter of rectifier equipment design. The electrical analog of the carrier decay was devised as a result of collaboration with F. W. Kelley, Jr., a consulting engineer with the Low Voltage Switchgear Department of the General Electric Company.

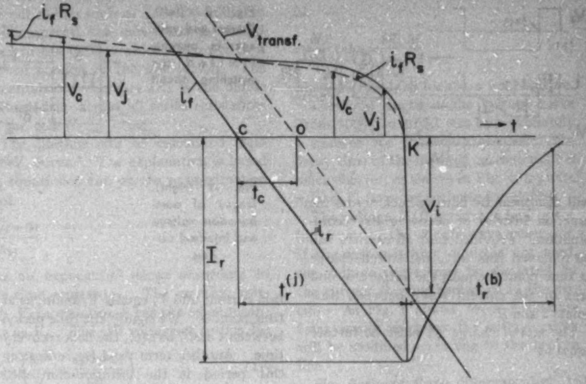


Fig. 1 (above). Voltage and current during commutation

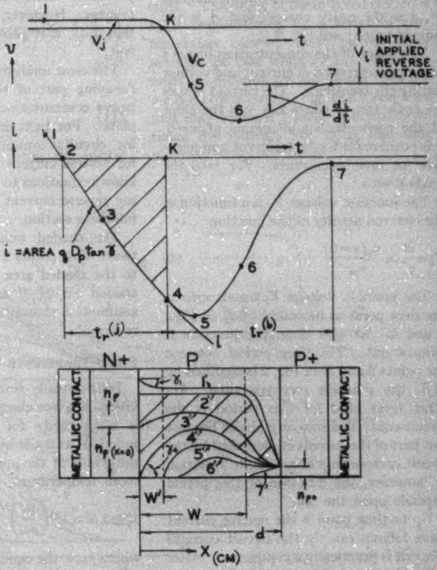


Fig. 2 (right). Cell voltage recovery current and concentration diagrams of injected carriers

ductance of the recovery current circuit:

$$\text{emf}_L = L\frac{di}{dt}$$

The current in a single-phase half-wave circuit during the reverse phase is

$$i_r = \frac{V_{tr}\text{ peak }\sin\omega t - V_j - L(di/dt)}{R} \tag{1}$$

where R and L are the ohmic resistance and the inductance of the circuit respectively. In the case where L is small and we neglect it,

$$i_r = \frac{V_{tr}\sin\omega t - V_j}{R} \tag{2}$$

When V_{tr} peak $\sin\omega t = -V_r$, i_r is equal to zero. This time point is indicated by C in Fig. 1. The current is leading the voltage. The phase shift is indicated by t_c in Fig. 1. Since

$$\sin\omega t_c = \frac{V_j}{V_{tr\text{ peak}}} \tag{3}$$

the phase shift is

$$t_c = \frac{\text{arc sin } V_j/V_{tr}\text{ peak}}{\omega} \tag{4}$$

or expressed by angle values

$$t_c = \frac{\text{arc sin } V_j/V_{tr}\text{ peak}}{\omega}\,\frac{2\pi}{360} \tag{5}$$

As a practical example, when $V_j = 0.6$ volts, $V_{tr} = 200$ volts, and $\omega = 2\pi\times 60$, then

$$t_c = \frac{\text{arc sin } 0.003\times 2\pi}{2\pi 60\times 360^\circ} = \frac{0.167^\circ\times 2\pi}{2\pi 60\times 360^\circ}$$
$$= 7.8\times 10^{-6}\text{ sec (seconds)}$$

This time is not the recovery time, but it is a phase shift between voltage and current occurring for pure ohmic load. The phase shift has a capacitive character and it is attributed to the so-called diffusion capacitance. For selenium cells $t_c = 0$. The cell voltage during the forward period (commonly called voltage drop) is divided into two parts: the junction voltage V_j and the voltage loss $I_f R_s$ of the semiconductor and soldering surfaces. The cell voltage curve remains continuous until it reaches the zero line. At time point C the current is zero; therefore the $I_f R_s$ component equals zero also. After time point C, the voltage loss $I_f R_s$ will be negative; see Fig. 1.

Carrier Concentration During Commutation

JUNCTION RECOVERY AND BULK RECOVERY

The forward conduction of a n^+-p junction is accomplished by injected minority carriers (electrons) on the p side. During the forward biased condition, the electron density is high on the p side. The concentration of the minority carriers are illustrated by the approximate diagrams in Fig. 2 for the case when the diffusion length L is much greater than the base width d.

The figure shows a typical current wave through the rectifier during commutation, the voltage across the cell (denoted by V_j in the forward phase and by V_c in the reverse phase), and it shows the approximate concentration diagrams corresponding to the marked time points of the current wave. It is assumed that the time duration of the recovery current is much less than the recombination time of the carriers. It is further assumed that all of the stored charge present at time point 2 will be recovered by the reverse current. This means that all injected electrons

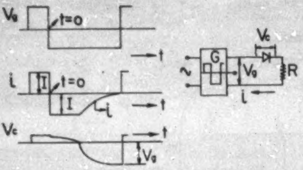


Fig. 3 (left). Current and voltage in current-limited d-c switching circuit

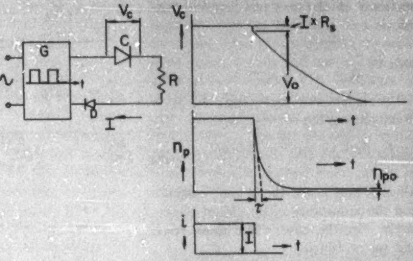


Fig. 4 (right). Decay of post-injection voltage and injected carriers

will disappear by flowing back across the junction instead of decaying by recombination. Provided that all carriers will be removed this way, the stored charge at time point 2 equals the ampere-second area of the current wave between time points 2 and 7.

The stored charge Q_s must be further equal to

$$Aq\int_0^d n_p(x)dx$$

where $n_p(x)$ is the concentration diagram 2′, q is the unity charge, and A is the junction area; see Fig. 2.

The shape of the concentration line is a common expression of current and voltage during the transient. The current equals $i=AqD_n \tan \gamma$, where $\tan \gamma$ is the slope of the concentration diagrams at $x=0$; γ is positive for forward current and negative for reverse currents. For $i=0$ the angle is zero.

The junction voltage V_j is a function of the electron density at the junction

$$V_j=\frac{kt}{q}\ln\frac{n_p(x=0)}{n_{p0}} \qquad (6)$$

The junction voltage V_j equals zero at the time point as indicated by K in Figs. 1 and 2. At this time, the junction is cleaned out. The time period between the points 2 and K in Fig. 2 is signified by $t_r(j)$, the junction recovery time. Another term used for this period is the circuit-controlled recovery phase, because this part of the current is controlled by the circuit components according to equation 1; however, the duration of this period depends upon the cell.

Up to time point 4 the reverse current wave follows exactly the circuit current. The cell is practically a conductor. After this point the cell assumes resistance.

The time point 5 is significant: at this time, the current slope $di/dt=0$ and therefore $L(di/dt)$ equals zero also. No voltage can occur across the inductance of the circuit, and the voltage V_c occurring across the cell is just equal to the applied voltage V_t.

In the time period between 5 and 7, the cell voltage V_c is larger than V_t, due to the induced voltage $L(di/dt)$, where L is the circuit inductance. At time point 7, the reverse current decreases to the saturation current and V_c equals V_t again, as at time point 5. We signify the time period between 4 and 7 as $t_r(b)$, the bulk recovery time. Another term used for indicating this period is the diffusion-controlled phase, because the current in this time period is determined by the diffusion process. However, for high commutation duty, it is also influenced by the circuit.

The clear analytical explanation of the decaying part of the reverse current in heavy commutation duty is not yet complete. For light commutation duty and for circuits containing series resistance for reverse current limitations, there are known equations to determine this decaying reverse current as given in the next following section.

The shaded ampere-second area between time points 2 and 4 is proportional to the shaded area between the concentration curves 2′ and 4′ (based on the assumption that all stored charges will be reversed).

FIELD TRANSIENTS

In the steady-state reverse-biased condition, a space charge region of width W is established; for abrupt p-n$^+$ silicon junctions as made by the alloying process, the width of the space charge region for room temperature

$$W\cong 3.57\times 10^3\left(\frac{V_r}{p_{p0}}\right)^{1/2}(\text{cm})$$

where p_{p0} is the equilibrium concentration of holes in the p region. For the situation given in Fig. 2 the applied voltage corresponds to the width W.

However, as discussed in the last paragraph, at time point 5 the total applied voltage occurs across the cell. The width of the space charge region cannot yet assume its steady-state value W because of the presence of electrons in the bulk. Its value is W'. This means a field transient in a part of the bulk during the commutation. The magnitude of this field transient depends upon the rate of change of voltage dv/dt between time points 4 and 5. The investigation as to whether or not this kind of partial field transient can hurt the cell is not yet complete.

LIFETIME OF MINORITY CARRIERS

At least two fundamental methods have been established for measuring the minority carrier lifetime. These tests determine the "effective" lifetime τ_{eff}, a measure for the volume and surface recombination. The equation for τ_{eff} is

$$\frac{1}{\tau_{eff}}=\frac{1}{\tau_v}+\frac{1}{\tau_s} \qquad (7)$$

where τ_v is the volume lifetime and τ_s is the surface lifetime. The term "lifetime τ" generally used in this report indicates the effective lifetime. We assume that the surface recombination is small, and the obtained value approaches τ_v. This lifetime may be (but not necessarily) equal to the crystal lifetime obtained by the light beam method. During production, the diode is subjected to heat treatment processes which may reduce the lifetime.

A method to determine the minority carrier lifetime has been described by Pell.[1] A square-wave voltage is applied to a diode as shown in Fig. 3. Both forward current and reverse current are limited by the resistance R. The reverse current is constant until the time when the concentration at the barrier will be equal to zero, after which it decays to the saturation current.

Both the straight-line portion and the decay portion of the reverse current are controlled by the lifetime and the diffusion equation

$$\frac{\partial n}{\partial t}+(n-n_{p0})/\tau=D_n\frac{\partial^2 n}{\partial x^2}$$

where n is the electron density, n_{p0} is the equilibrium density of electrons in the p region, and D_n is the diffusion constant for electrons. The solution of the diffusion equation as a time function gives the equation of the reverse current during the decay time

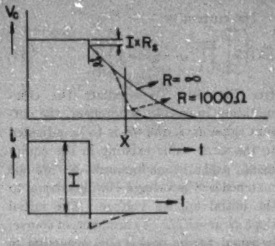


Fig. 5. Postinjection voltages in circuits of very high and moderately high resistance

$$i=\left(\frac{D_n n_{p0}}{L_n}\right)-\frac{D_n(n_{x=0})}{L_n}\times\left(\frac{\exp -t/\tau}{\sqrt{t/\tau}}-\operatorname{erfc}(t/\tau)^{\frac{1}{2}}\right) \qquad (8)$$

where $n_{x=0}$ is the electron density at $x=0$, D_n is the diffusion length, and

$$\operatorname{erf}x=1-\left(\frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt\right) \qquad (9)$$

This equation holds true for small injections and a junction with $d>L_n$. The lifetime was calculated from the equation and experimentally obtained decay current curves.

The time duration $t_r^{(f)}$ of the straight-line portion of the reverse current is related to the lifetime by[2]

$$\operatorname{erf}\sqrt{t_r^{(f)}/\tau}=\frac{1}{1+(I_{rev}/I)} \qquad (10)$$

Another method of determining the lifetime is to inject minority carriers into the bulk by means of a forward current pulse which is much longer than the lifetime, and to observe the cell voltage V_c after the pulse is terminated. For low current values the current pulsing can be established by a square-wave generator; see Fig. 4. The lifetime of the cell C will be measured. The square-wave generator as illustrated in Fig. 4 produces forward current pulses of amplitude I determined by the resistor R and the internal resistance of the generator. The rise time of the generator output should be 0.2 μsec (microsec) or less.

After the current pulse is terminated, the cell C will for a short time be the source of voltage in the circuit. The auxiliary diode D, connected in series with the cell C, must be a fast-switching type of rectifier to prevent a reverse current flow. Its recovery time should be one order of magnitude shorter than that of the cell C. By means of this diode an open-circuit condition is established after the current pulse ends.

The voltage curve V_c shows the voltage generated by the cell. For low forward current values such as 1 to 100 milliamp (milliamperes) the curve has an initial straight-line portion which is followed by a curved portion in the low region. The instantaneous decrease IR_s is the ohmic voltage drop across all series resistances of the cell C.

The lifetime can be calculated from the V_c curve.[3] The explanation is based on equation 6 and on the approximation that

$$\frac{n_p(t=0)}{n_{p0}}$$

has an exponential decay according to the time constant τ. The junction voltage V_j in equation 6 has the same meaning as V_c. The cell voltage at $t=0$ is V_0:

$$V_0=\left(\frac{kT}{q}\ln\frac{n_p(t=0)}{n_{p0}}\right) \qquad (11)$$

and

$$V_c(t)=\frac{kT}{q}\ln\frac{n_p(t=0)}{n_{p0}}\exp-\frac{t}{\tau}=V_0-\left(\frac{kT}{q}\frac{t}{\tau}\right) \qquad (12)$$

The lifetime

$$\tau=\frac{kT}{q}\left(\frac{t}{V_0-V_c(t)}\right) \qquad (13)$$

The τ is calculated from the initial slope α, obtained for the first short period Δt

$$\tau=\frac{kT}{q}\tan\alpha=\frac{kT}{q}\frac{\Delta t}{\Delta V} \qquad (14)$$

where $kT/q=1/38$ for room temperature.

The total length of the voltage curve (from 100% to 10%) was approximately 1 millisec for the cell measured. The decay time of the injected carriers $n_p(t)$ is much shorter, as shown in Fig. 4, because n_p and V_c have logarithmic correlation.

The voltage $V_c(t)$ is independent of the frequency of the square-wave generator. If the fast-switching diode is not used, the voltage curve may be distorted by the current flowing backward in the circuit. As Fig. 5 shows, for a 1,000-ohm total circuit resistance the voltage curve will be modified according to the broken line.

Fig. 6 shows three different kinds of clean-out processes in low-voltage circuits. The decay of the injected electrons n_p is demonstrated by a series of concentration curves; these curves characterize the case when $d>L_n$. The time units are different for each case. For the a-c d-c power circuit, Fig. 6(A), one time unit may indicate 1 μsec. For unidirectional d-c pulsing, Fig. 6(C), one time unit may indicate 200 μsec.

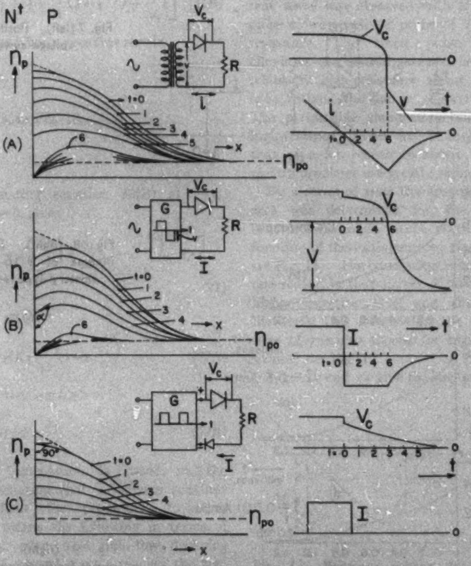


Fig. 6. Concentration diagrams of injected carriers

TESTING OF HIGH-LEVEL LIFETIME

The output of electronic square-wave generators is only approximately 100 milliamp. For testing the lifetime at higher currents the circuit in Fig. 7 was used. The switch S interrupts the direct current and the oscilloscope shows a single voltage decay curve. The advantage of this circuit is its simplicity, but one disadvantage is that for really high current an arcing occurs and precludes accurate measurements. Other circuits which did not produce arcing, were also used to measure lifetime.

$V_c(t)$ curves obtained by means of the interruption circuit are given in Fig. 7. A constant value of τ corresponds to a straight line. The change of the slope indicates the change in lifetime. The lifetime as a function of direct current is given in Fig. 8. The injection level is given by $V = n_{pi}/n_0 + p_0$.[4] $V > 1$ indicates high injection while $V < 1$ indicates low injection; n_0 and p_0 indicate the equilibrium values in the p region.

Commutation in A-C D-C Rectifier Circuits

TESTING CIRCUIT

In common a-c d-c circuits the current does not change abruptly during the transition from the forward to the reverse direction. It follows in most cases a sine or cosine line and has a slope at $i = 0$, determined by the circuit. Since the cell represents a conductor for a short time after $i = 0$, the slope does not change at this time and the current wave as it passes through zero is a straight line. For 60 cps (cycles per second) or even for 600 cps frequency the recovery process is a very short portion of the cycle. The current line may therefore be considered as an unbroken straight line during the commutation.

The circuit of the recovery path has relatively high inductance and low resistance as opposed to the high resistance, low inductance circuit of Fig. 3.

The investigation of recovery current in large installations is impractical, particularly when we want to test under overload or short-circuit conditions. These problems are avoided by using the circuit in Fig. 9, which is a pulsing circuit producing a single commutation process. The pulsing is initiated by the mercury switch M. Prior to the switching, a steady-state forward current will be reduced to zero and then will reverse along a straight-line curve in a manner similar to that in practical power circuits. For a short time the current is controlled by the circuit only, because the cell is a conductor.

The current is

$$i = I_f + \left[\frac{V}{R_1}\left(\exp\frac{-tR_1}{L} - 1\right)\right] \tag{15}$$

(see Fig. 9). To simulate the commutation in rectifier equipment, the current slope di/dt at $i = 0$ is to be adjusted to the same value existing in the equipment, and the inductance has to be matched. The voltage V will be equal to the initial applied voltage. The initial slope $di/dt = V/L$. I_f need not, of course, be equal to the peak value occurring in the equipment. The instantaneous current values occurring at time 3τ (or longer) before $i = 0$ have no influence on the stored energy in the cell at the time when $i = 0$.

INJECTION LEVEL PRODUCED BY LONG STRAIGHT-LINE CURRENT DECAY

The recovery current is attributed to a stored charge existing at the time when $i = 0$. The question therefore arises as to what the stored charge is or, in other words, what the concentration of injected carriers is when a decaying straight-line current wave reaches zero. This problem was analyzed by means of an R-C analog, as shown in Fig. 10.

We assume that when a forward current through a rectifier is reduced abruptly to zero, the number of injected

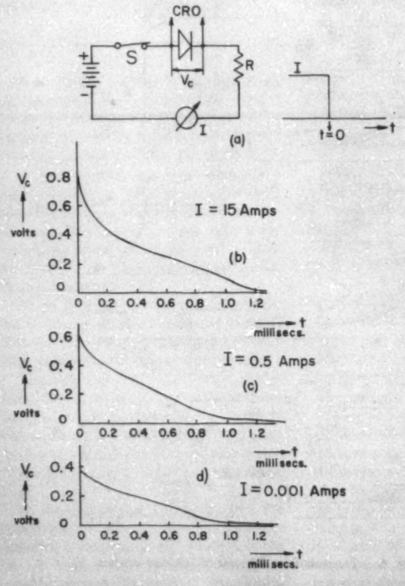


Fig. 7 (left). Postinjection voltage curves

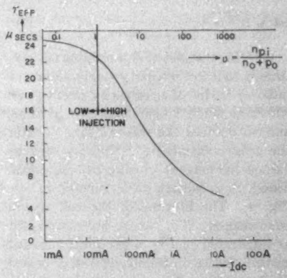


Fig. 8 (right). Effective lifetime of 9/16-inch-diameter p-type rectifier

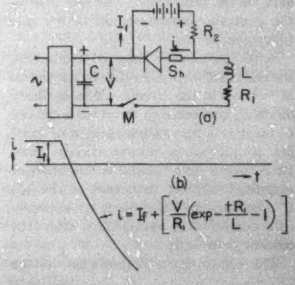


Fig. 9 (right). Pulsing circuit for single commutation pulse

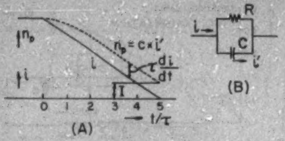


Fig. 10. A—Injected carriers during current decay. B—Equivalent circuit

carriers n_p decays exponentially according to the time constant τ. In the analog circuit i represents the current through the rectifier; i_c' is the current through capacitor C; and $i_r' = i - i_c'$, the current through the resistor. The time constant $\tau = RC$.

Let n_p = number of carriers $= Ki_r'$. From circuit theory

$$Ri_r' = \frac{Q_0}{C} + \frac{1}{C}\int i_c' dt \tag{16}$$

Using the Laplace transformation yields

$$I'(s) = I_r'(s) + I_c'(s) \tag{17}$$

$$RI_r(s) = \frac{Q}{Cs} + \frac{I_c'(s)}{Cs} \tag{18}$$

Combining equations 17 and 18 gives

$$I'(s) = \frac{Q_0}{\tau s} + I_c'(s)\left(1 + \frac{1}{\tau s}\right) \tag{19}$$

or

$$I_c'(s) = \frac{sI'(s) - Q_0/\tau}{s + 1/\tau} \tag{19A}$$

Let us assume that $i = I_f$ during the time period between $t = -100\ \tau$ and $t = 0-$. By inspection of the R-C analog

at $t = 0-$, $i = i_r' = I_f$

at $t = 0$, $I_f R = \frac{Q_0}{C}$

at $t = 0$, $Q_0 = I_f RC = \tau I_f$

Equation 19(A) may be expressed as

$$I_c'(s) = \frac{sI'(s) - I'f}{s + 1/\tau} \tag{20}$$

During the decaying time of the current wave

$$i_{(t \gtreqless 0)} = I_f' - \frac{\Delta i}{\Delta t} t \tag{21}$$

Using the Laplace transformation gives

$$I'(s)_{(t=0)} = \frac{I_f}{s} - \frac{1}{s^2}\frac{I}{t} \tag{22}$$

Equations 20 and 22 may be combined, yielding

$$I_{c(s)}'{}_{t \gtreqless 0} = \frac{-\Delta_i/\Delta_t}{s(s + 1/\tau)} \tag{23}$$

The inverse transform of equation 23 is

$$i_c' = -\tau \frac{\Delta i}{\Delta t}(1 - e^{-t/\tau}) \tag{24}$$

The number of carriers $n_p = Ki_r'$ may be expressed by equation 16:

$$n_p = K\left(\frac{Q_0}{\tau} + \frac{1}{\tau}\int i_c' dt\right) \tag{25}$$

or

$$n_p = K\left(\frac{Q_0}{T} - \frac{\Delta i}{\Delta t}\int_0^t \left(1 - e^{-\frac{t}{\tau}}\right) dt\right)$$

$$= K\frac{Q_0}{\tau} - K\frac{\Delta_i}{\Delta t}\left(t + \tau\left(e^{-\frac{t}{\tau}}\right) - \tau\right) \tag{25A}$$

or

$$N_p = K\left[\frac{Q_0}{\tau} + \tau\frac{\Delta_i}{\Delta t}\left(1 - e^{-\frac{t}{\tau}}\right) - \frac{\Delta_i}{\Delta t} t\right] \tag{25B}$$

The time t_1 when i reaches zero is

$$t_1 = \frac{\Delta_t}{\Delta t} I_f$$

which is, in Fig. 10,

$$\frac{t_1}{\tau} = 5 \tag{26}$$

Evaluating equation 25(B) at $t = t_1$, when $i = 0$, gives

$$n_{p(i=0)} = K\left[\frac{Q_0}{\tau} + \tau\frac{\Delta_i}{\Delta t}\left(1 - e^{\frac{-t_1}{\tau}}\right) - I_f.\right] \tag{27}$$

Since $I_f = Q_0/\tau$

$$n_{p(i=0)} = K\tau\frac{\Delta_i}{\Delta t}\left(1 - e^{\frac{-t_1}{\tau}}\right) \tag{27A}$$

and for the case $t_1 >> \tau$

$$n_p(i=0) = K\tau\frac{\Delta_i}{\Delta t} \tag{27B}$$

This calculation is based on the approximation that τ remains constant throughout the procedure. Equation 27(B) means the following in terms of current: when a straight-line decaying forward current wave i reaches the zero line, the stored energy at $i = 0$ is equal to the stored energy caused by a direct current $I = \tau \times di/dt$; see Fig. 10.

In other words, the injection level of a straight-line decay current wave i, at the time when $i = 0$, equals $\tau \times di/dt$.

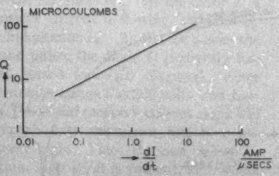


Fig. 11. Q versus dI/dt for 9/16-inch-diameter p-type rectifier

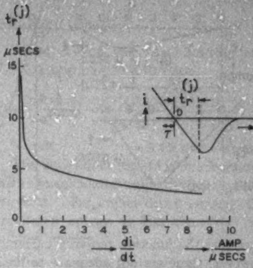


Fig. 12. Junction recovery time versus dI/dt at I=0

Recovery Current in A-C D-C Circuits: Destructive Oscillation

Recovery currents obtained by the pulsing circuit given in Fig. 9 were analyzed. To simulate the commutation of a common rectifying circuit, the forward current was adjusted as a straight line of slope di/dt, having no break within 3τ before reaching the zero line. The current wave was observed with a current pulse viewing resistor probe of 0.25-ohm resistance. The single pulse greatly alleviates the shunt problem, because a relatively high-resistance shunt can be used without the danger of overheating. The permissible shunt inductance of a high-resistance shunt may be higher than that of a low-resistance shunt and the pickup problems are greatly reduced.

By a series of tests the ampere-second area was determined for the typical 9/16-inch-diameter p-type rectifier as a function of the initial current slope di/dt; see Fig. 11. The correlation is not linear; one reason is that higher di/dt means a higher injection level and thus lower lifetime. This means that the recovery time $t_r(j)$ is also shorter for higher di/dt values; see Fig. 12.

Figs. 11 and 12 give the means to pre-

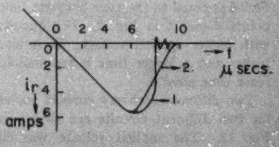


Fig. 13. Recovery currents for different values of inductance

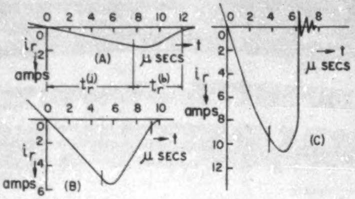


Fig. 14 (left). Recovery currents for different di/dt with L = constant

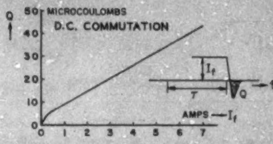


Fig. 15 (right). Ampere-seconds versus forward curren

determine the junction recovery current. However, these are not adequate to predetermine the bulk recovery current, because it is also influenced by the circuit parameters. Equation 8 indicates that the decaying portion of the recovery current is independent of voltage. The general equation for the decaying current in rectifiers with $d<L$,

$$i = K_1 + K_2 \exp -(D\pi^2 t/d^2)$$

shows also that the current is controlled only by diffusion; see reference 1. However, for both p^+-n-n and n^+-p-p alloyed junctions beyond a certain applied voltage, a voltage dependence is present.

The reverse voltage across the cell, V_c during the bulk recovery period, equals the sum of the initial applied voltage and the voltage Ldi/dt, induced by the circuit inductance. The distortion of the current wave attributed to voltage influence is therefore larger in highly inductive circuits. The current wave defined by the above-mentioned equations starts to decay immediately at $t = t_r(j)$, at the time when the reverse voltage appears.

In power rectifier equipments, and likewise in the circuit in Fig. 9, where the commutation circuit contains relatively high inductance and low resistance, the current follows the pattern of Fig. 2, where the current keeps increasing after $t = t_r(j)$. In other words, the current wave does not start to decay at time point 4, but keeps increasing and reaches the peak value at point 5, where the total applied voltage occurs across the cell. At time point 5 the resistance of the cell equals $R_b = V_{\text{applied}}/I_{\text{reverse}}$ peak.

A final time period is required for the cell resistance to reach R_b. This time period is equal to the time between time points 4 and 5 in Fig. 2. For circuits with higher applied voltage, R_b is also higher and a longer time is required to reach time point 5.

Two different recovery current waves for two different circuits are shown in Fig. 13. The applied voltage was 60 volts in case 1 and 15 volts in case 2. The circuit inductance was 15 and 60 μh (μhenrys) respectively; thus the initial di/dt was equal to 1 amp/μsec (amp per μsec) in both cases, and hence the junction recovery currents also are equal. However, the bulk recovery currents are different.

As the current reaches its peak, the resistance of the cell R_b is equal to 15 volts divided by 6 amp = 2.5 ohms for case 2 and 60 volts divided by 6 amp = 10 ohms for case 1. As Fig. 13 shows, it took a longer time to reach the peak in case 1, representing the high-inductance high-voltage circuit.

The ampere-second area of the recovery current must be the same for both cases. Since the bulk recovery current in case 1 decreases too slowly, the stored energy will be cleaned out in a relatively short time and the current ends in an abrupt decay. This high rate of change of the current results in a damped oscillation of frequency

$$f = \frac{1}{2\pi}\frac{1}{\sqrt{LC_{\text{cell}}}}$$

The value of the first peak is equal to $V_{\text{peak}} = \Delta I\sqrt{L/C}$, where L is the circuit inductance, C is the cell capacitance, and ΔI is the abrupt current; ΔI is 8 amp in Fig. 14(C).

In the test circuit using only one single cell and inductance $L = 20$ μh, the frequency of the oscillation was between 1 and 10 megacycles. The rise time of the first voltage spike may be less than the rise time of the conventional oscilloscope which shows a voltage lower than the real value. The Tektronix *535* oscilloscope showed approximately 50% of the value showed by the Tektronix *545A* in one particular case.

If the voltage spike exceeds the cell voltage rating, the cell may break down. Cells of normal geometry show a crater on the edge of the gold electrode after this kind of commutation breakdown. The highest field apparently exists across this edge.

The occurrence of the destructive oscillation is shown in Fig. 14 for $L = 15$ μh. The initial $di/dt = V/L$ is controlled by adjustment of V. Above a certain critical di/dt the abrupt current change occurs.

D-C Commutation

CIRCUITS PRODUCING LOW I, HIGH DI/DT COMMUTATION

We can distinguish between two typical types of commutation. The forward current corresponding to the current waves in Figs. 13 and 14 represents a long line, compared with the lifetime τ, and has a constant slope, which, occurring in common power rectifiers, is in the range of 0.1 to 1.0 amp/μsec for normal conditions. This represents the case of a-c d-c commutation.

Besides these common cases, commutation types exist as shown in Fig. 15, where a steady-state direct current flows prior to the commutation. The rate of change of the current in this case is also equal to V/L. With proper adjustment, the test circuit in Fig. 9 produces this type of commutation which represents d-c commutation. When the current slope V/L is high, the reversal of the current is carried out in a much shorter time than τ; see Fig. 15. Thus we may assume that the stored energy at the time that $i = 0$ corresponds to I_f.

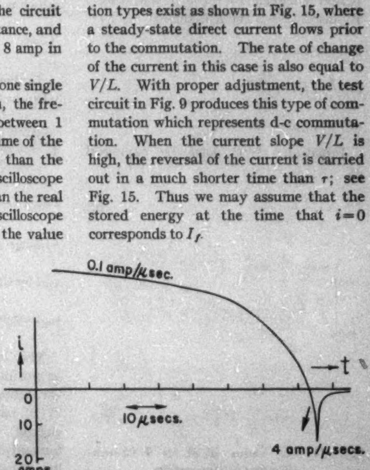


Fig. 16. Individual cell current in a parallel circuit

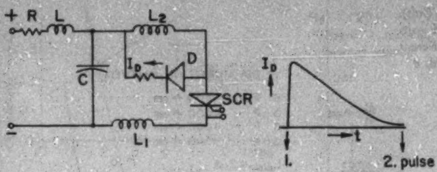


Fig. 17. Ignitron firing circuit

In a-c d-c commutation as treated in the previous section, the stored energy was proportional to the current slope and high di/dt caused high stored energy. In the d-c case the stored energy is proportional to the current I_f and very low I_f may be combined with very high current slope. This represents the case of "low injection, high di/dt." This case may result in an undesirable modification of the recovery current waveform, and may aggravate the commutation duty.

Two practical cases representing this type of commutation are given in Figs. 16 and 17. Fig. 16 shows the current flowing through a cell installed in a leg of a power rectifier, installation having 36 cells in parallel. The cell having the highest inductance assumes the full slope of the total current in that leg as the current approaches zero. Thus, we have the low injection, high di/dt case.

The current flowing prior to the commutation is not, of course, a pure direct current; still, the commutation as shown in Fig. 15 is approached. Because of space limitation the detailed analysis of this commutation (which might lead to breakdown when no protection is used) is not presented here.

Fig. 17 shows a newly developed ignitron excitation circuit which was designed by C. E. Rettig, where the rectifier cell is undergoing the same type of commutation, and a protective capacitor is required in order to avoid instantaneous breakdown. The scr (silicon-controlled rectifier) initiates two successive current pulses through L_2 (primary of a pulse transformer), by discharging C. After the first pulse, the L_1, L_2, C oscillatory circuit develops a current reversal and the scr turns off. In the diode D a transient current is flowing, supplied by the inductance L_2. As the scr passes the second pulse, the diode D is slightly forward biased (see $I_d(t)$ diagram). Since the voltage across C is 200 volts, and $L_1 = 2\mu h$, the initial recovery current slope will approach 100 amps/μsec, which is extremely high. Thus, the case of low injection, high di/dt is present, and would cause breakdown if the protective capacitor were not used.

Commutations of this kind are to be expected in many circuits, where scr (or other switching devices) are associated with rectifiers.

Stored Energy During Commutation From Direct Current

The diagram of the ampere-second area of recovery currents versus I_f forward current are given in Fig. 15. The decay time of I_f was equal to approximately 100 milliμsecs, which is much shorter than the lifetime.

The stored energy for the forward current I_f and for 1-cm^2 junction area, and for $L > d$ is

$$Q_s = dn_{p1}q = \frac{I_f \times d \times n_{p0}}{I_s} q \qquad (28)$$

where I_s is the saturation current of 1-cm^2 junction area.

$$I_s = \frac{qn_{p0}D_n}{L_n} = \frac{qn_{p0}\sqrt{D_n}}{\sqrt{\tau_0}} \text{ amp} \qquad (29)$$

Thus, for the A cm^2 area

$$Q_s = \frac{AdI_f\sqrt{\tau_0}}{\sqrt{D_n}} = AI_f\tau_0 \frac{d}{L} \text{ amp-sec} \qquad (30)$$

For the case $L_n = d$

$$Q_s \approx \frac{AI_f\tau_0}{2} \text{ amp-sec} \qquad (31)$$

where τ_0 is an extremely low-level lifetime corresponding to the saturation current I_s.

Numerical Sample

When $A = 1.6$ cm^2, $d = 1.3 \times 10^{-3}$ cm, $I_f = 0.4$ amp, $\tau_0 = 20 \times 10^{-6}$ sec

$$Q_s = \frac{1.6 \times 1.3 \times 10^{-3} \times 0.4 \times \sqrt{20 \times 10^{-6}}}{\sqrt{33}}$$

$$= 6.4 \times 10^{-6} \text{ coulombs}$$

The diagram in Fig. 15 is in accordance with the calculated values for small direct currents. For higher direct currents the recovered energy is less than the calculated energy.

Recovery Current Diagrams for Different D-C Injections

A family of diagrams of i_r versus t for various direct currents I is shown in Fig. 18. The reversal of the forward current was carried out in a short time compared with the lifetime. The curves were observed in the circuit shown in Fig. 9. The circuit inductance $L = 2$ μh, the applied voltage $V = 64$ volts, the initial slope $di/dt = V/L = 32$ amp/μsec.

For forward currents smaller than 0.25 amp, abrupt current changes and excessive voltage spikes of 300 to 500 volts occurred. By increasing V to 80 volts, the abrupt current change and voltage spike occurred even for a 0.4-amp forward current. The circuit inductance $L = 2$ μh is a very low value; however, it represents a practical case. When the transformer inductance is eliminated by parallel capacitors in power rectifier installations, the remaining inductance of the bus system may be only a few microhenrys.

Influence of Cell Parameters on Bulk Recovery Currents

Certain cell parameters, which may be partly controlled during fabrication, have influence on the bulk recovery current.

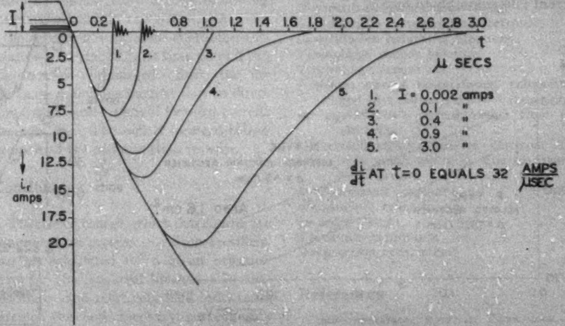


Fig. 18. Recovery currents for different injections

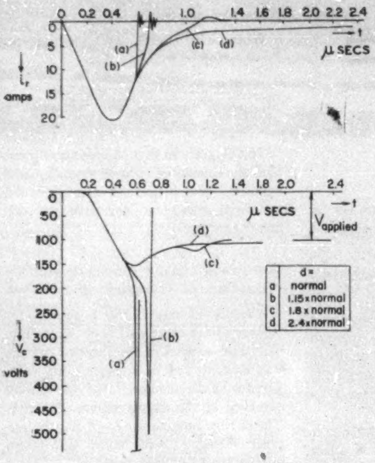


Fig. 19 (left). Current and voltage for different base width d

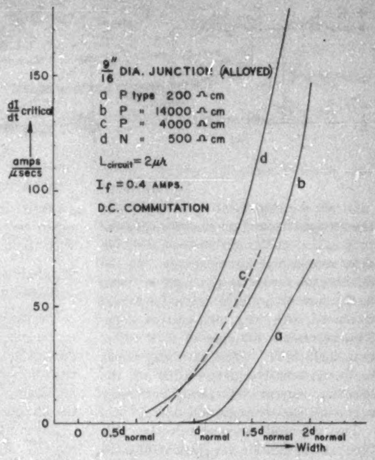


Fig. 20 (right). Critical di/dt versus base width d

Such a parameter is the width *d* of the p region in the n^+-p junction (or of the n region in the p^+-n junction). To investigate this relation, a series of cells of different parameters and crystal properties were produced and tested. Fig. 19 shows the influence of one parameter, the width *d* of the p region in the n^+-p junction. The junction recovery current and the first portion of the bulk recovery current are identical. The effect of the width appears in the last part of the bulk recovery current. The ampere-second area increases with increasing width *d*. The area increase was expected to be proportional to the increase of *d*, but it is slightly less than proportional. The area increase results in an effective prolongation of the bulk recovery time and, instead of an abruptly decaying wave, the current falls smoothly to zero.

The voltage curves at the bottom of Fig. 19 show the developing of the spikes corresponding to the current curves above. For the sake of clarity, the oscillations following the spikes are not illustrated on the curves. The peak value of (*a*) was 1,000 volts.

Adjustment of the test circuit, in Fig. 9, was $I_f = 0.4$ amp, $V = 130$ volt, $L = 2$ μh. Current slope $di/dt = 130/2 = 65$ amps/μsec.

Fig. 20 shows a comprehensive illustration of the influence of the width *d*. The critical *di/dt* indicates the slope at which the abrupt current change and voltage spike occurs. These sheets may be obtained for any circuit inductance and commutation type (a-c–d-c or d-c). This present sheet was obtained for $L =$ 2 μh inductance and for d-c commutation with 0.4 amp forward current. For higher inductance the critical slope is much lower. For instance, with $L =$ 20 μh, for the curve *C*, and normal *d*, the critical current slope equals approximately 1 amp/μsec, instead of 20 amp/μsec. The value of d_{normal} was equal to 0.5 L_n.

For a given *V*, the highest voltage spike occurs for a certain *L*, in most cases between 20 and 100 μh. Above and below these values of inductance the spike is smaller.

When the parameters of a rectifier circuit are such that *di/dt* cannot exceed the critical *di/dt* under the given overload or short-circuit conditions, no protective capacitors are necessary.

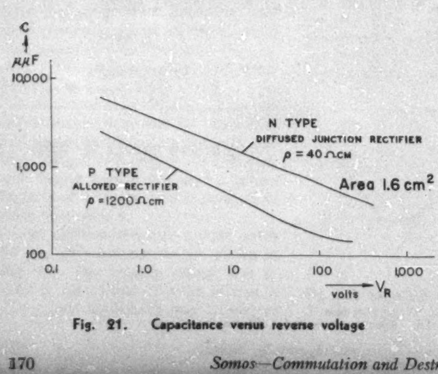


Fig. 21. Capacitance versus reverse voltage

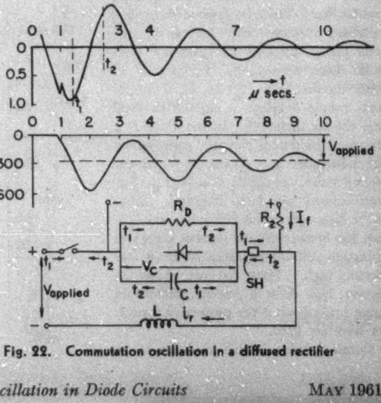


Fig. 22. Commutation oscillation in a diffused rectifier

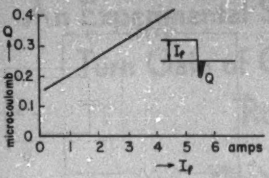


Fig. 23 (left). Ampere - seconds forward current

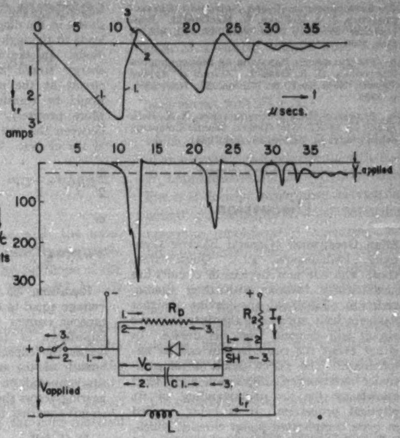


Fig. 24 (right). Repetitive ringing

Commutation of Diffused Junction Rectifiers

The commutation of diffused junction rectifiers may differ substantially from the commutation in alloyed junction types. The diffused junction has comparatively very low lifetime and very high capacitance. The low lifetime results in a small recovery current. The high capacitance results in a large oscillating current. Thus, the recovery current may be overwhelmed by the oscillating current.

The cell capacitance can be varied by production techniques and it depends also on the crystal resistivity. Fig. 21 shows the comparison between the capacitance of a normal p-type alloyed junction and a low-resistivity diffused junction. ($\zeta=50$ ohm-cm, n type).

The commutation process of the low resistivity diffused n-type junction rectifier resembles the switching transient oscillation of an inductance-capacitance-resistance series circuit, where the peak value of the voltage across the capacitors is equal to or less than twice the applied direct voltage; see Fig. 22. The frequency equals $1/(2\pi\times\sqrt{LC})$. Thus, the capacitance of these rectifiers acts as a built-in protection as in the selenium rectifier.

The current wave $i_r(t)$ in Fig. 22 may lead to the conclusion that forward current is flowing through the cell at $t=t_3$, but this is not the case. The equivalent circuit in Fig. 22 shows the directions current flow (conventional) through the cell and shunt at $t=t_1$ and $t=t_2$. The rectifier cell composed of R_D and C, is shown connected into the test circuit in Fig. 9.

At $t=t_1$ the capacitor is charged. At $t=t_2$ the capacitor discharges and it is the source of energy having a higher emf than the power supply. During this discharging, the current flows backward in the circuit and shunt, as shown by the arrows. However, the current through R_D, the reverse leakage, does not change its direction. At $t=t_3$ the reverse leakage does not flow through the shunt at all; it is supplied by C.

The commutation in Fig. 22 was obtained with 5-amp forward current and $L=150$ μh, $V=280$ volts.

For the above-mentioned diffused junction rectifier the ampere-second area versus forward current I_f for the case of d-c commutation is shown in Fig. 23. Comparison with Fig. 15 shows the substantial difference due to differences in lifetime.

Repetitive Oscillation

Under certain conditions, mainly in circuits with high inductance, the commutation process will be repeated many times. This is analyzed in Fig. 24, showing also the equivalent circuit of the cell connected into the test circuit; see Fig. 9. The curves were obtained with $I_f=0.4$ amp, $L=75$ μh, $V=25$ volts.

The bulk recovery current is prolonged by the charging current of the cell capacitance C. At time point 2 the capacitor starts to discharge by a current flowing counterclockwise in the circuit. At time 2 the current has the same characteristics as at $t=t_2$ in Fig. 22. At time point 3 the capacitor is discharged. Still, the inductance maintains current flow in direction 3 and causes a small forward current through R_D. The cell is forward-biased again, and the commutation repeats.

Conclusions

The stored charge, which determines the commutation process of silicon rectifiers, can be calculated for a given injection level by the low-level lifetime. In low-resistance and relatively high-inductance circuits the bulk recovery current may assume various forms for the same stored charge. During the bulk recovery a transient field increase occurs along a portion of the bulk. Beyond a critical di/dt an abrupt termination of the bulk recovery current may cause oscillation and excessive voltage spikes. In this case protective capacitor circuits are necessary to assure trouble-free operation. Rectifiers combining high cell capacitance with low lifetime produce a harmless oscillation having a maximum peak voltage equal to twice the initial applied voltage.

Nomenclature

n_0 = equilibrium density of electrons
p_0 = equilibrium density of holes
n_p = density of electrons in p region
n_{pi} = density of injected electrons in p region
n_{p0} = equilibrium density of electrons in p region
p_{p0} = equilibrium density of holes in p region
$n_{(x=0)}$ = density of electrons at the barrier
d = width of p region of n$^+$-p cells
W = width of space charge region
D_n = diffusion constant for electrons
L_n = diffusion length for electrons
V_j = voltage across junction
V_c = voltage across cell
V_0 = open circuit postinjection voltage at end of a forward current pulse
$\tau=\tau_{eff}$ = effective recombination time for electrons
τ_0 = recombination time of electrons for very small injection, due to saturation current
k = Boltzmann constant
T = absolute temperature
q = unity charge
L = circuit inductance
A = junction area in cm^2

References

1. RECOMBINATION RATE IN GERMANIUM BY OBSERVATION OF PULSED REVERSE CHARACTERISTICS, E. M. Pell. *The Physical Review*, New York, N. Y., vol. 90, no. 2, Apr. 15, 1953.

2. SEMICONDUCTOR DIODE SWITCHING CHARACTERISTICS, J. R. Madigan, W. MacDonald. *Semiconductor Products*, New York, N. Y., vol. 3, Feb. 1960.

3. ON TRANSIENT BEHAVIOR OF SEMICONDUCTOR RECTIFIERS, B. R. Gossick. *Journal of Applied Physics*, New York, N. Y., vol. 26, Nov. 1955, p. 1356.

4. ELECTRON-HOLE RECOMBINATION, R. N. Hall, *Report no. 55-RL-1316*, General Electric Company, Schenectady, N. Y., vol. 58, July 1955.

◆

Discussion

Allan Greenwood (General Electric Company, Philadelphia, Pa.): Quite often, those who use new devices in circuits are insufficiently familiar with their characteristics. Likewise, the device designer may not fully appreciate the duty that the circuit will demand from his device. In either event the consequences can be unfortunate. The silicon diode, a comparatively new device, requires a specialized knowledge for an understanding of its physical properties. It is being applied in quite complicated power circuits, which, one might almost say, are in a perpetual state of transience. To appreciate the behavior of circuits under such conditions again calls for special competence.

Here is a situation with ample latitude for error. The author of this paper has made a most useful contribution by showing how, in the complicated process of commutation, the characteristics of the diode and the circuit interact. A particularly interesting manifestation of this interaction, which he points out, is the excitation of high-frequency oscillatory voltages across the cell, which can reach destructive proportions.

Oscillations similar to those described occur in other areas of power system engineering. The basic stimulus in each instance is a sudden suppression of current. A well-known example is the abrupt quenching of the arc in a circuit breaker, which causes the current to come to zero prematurely, ahead of its natural zero point. This phenomenon, usually referred to as current chopping, can also generate excessive overvoltages. With the diode, the suppression is brought about after the current reverses. The reverse current represents a sweeping out of the carriers from the junction causing the diode to block. At the moment of blocking the circuit might be represented as shown in Fig. 25.

The current is shown flowing momentarily in the reverse direction through diode 1. When the diode blocks, the current is forced into the capacitance of the device, shown in the diagram as C; it is this which leads to the high inverse voltage across the diode. In the absence of damping, the energy stored in the inductance of the circuit at the moment the diode blocks would be transferred to the capacitance. More precisely, the energy will oscillate between the two at the natural frequency of the circuit. We may write

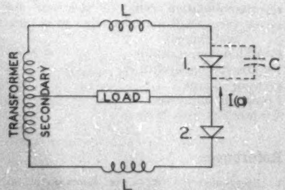


Fig. 25. Circuit at moment of blocking

$$\frac{1}{2}LI(0)^2=\frac{1}{2}CV^2$$

or

$$V=I(0)\sqrt{\frac{L}{C}} \qquad (32)$$

Equation 32 indicates that a peak voltage equal to the product of the instantaneous current suppressed and the surge impedance of the circuit would appear across the capacitance as a direct consequence of the suppression. It would be superimposed on the normal voltage appearing across the cell as it recovers. The circuit equation, neglecting damping, will be

$$V_T=LC\frac{d^2V}{dt^2}+V \qquad (33)$$

where V is the voltage across the capacitor (i.e., across the diode) and V_T is the instantaneous transformer voltage. Entering the initial conditions, namely, that at time zero the current is I (0) and the capacitor is uncharged, the following solution for V emerges:

$$V=V_T(1-\cos\omega_0t)+I(0)\sqrt{\frac{L}{C}}\sin\omega_0t \qquad (34)$$

where $\omega_0^2=1/LC$.

V_T is treated as constant in equation 34. This approximation is justified in view of the negligible supply frequency variation in the transformer voltage during the brief period of the transient. The first term in equation 34 is the normal recovery transient; the second is a direct consequence of the current suppression. In the event that the circuit has a high surge impedance, a small suppression can generate a considerable voltage, quite enough to destroy the cell. The beneficial effect of adding capacitance across the cell, to reduce the surge impedance, is obvious.

Equation 34 is, of course, an approximation. Damping is neglected and C is treated as a constant although it is well known that the capacitance of a cell varies considerably during the transition from the conducting to the blocking state. Then again, the diode would conduct once more during the oscillation were it to see a forward voltage. Equation 34 indicates that this will occur. We might expect this to lead to the "repetitive ringing" referred to in the paper.

Istvan Somos: Dr. Greenwood's analysis explains explicitly the process taking place upon the abrupt suppression of the reverse recovery current, which occurs when critical circuit parameters are exceeded. The surge impedance $\sqrt{L/C}$ of diodes treated in the article, having 1.6 cm² junction area, in a circuit of 10-μh inductance, is calculated to be 260 ohms, based on 150-μμfarad capacitance. Owing to the damping, the actual value is smaller; it was found by testing to be approximately 200 ohms. This means that in a circuit with a 10-μh inductance the first peak value of the voltage oscillation equals the chopped current multiplied by 200.

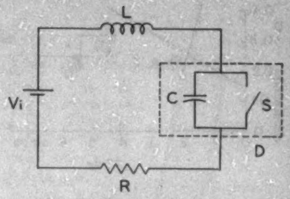


Fig. 26. Circuit conditions at instant of current chopping

An equation similar to the derived equation in the discussion may also be obtained for the switching circuit, which is similar to the "low-tension ignition circuit" of gasoline engines when the spark is generated across the switch as the switch located within the cylinder is opened; see Fig. 26.

The diode D is represented by the capacitor C and switch S. The circuit represents the conditions at the instant of the current chopping. V_i is analogous to the initial applied voltage denoted by V_i in the article.

The current chopping is accomplished by opening the switch S. If this time is considered as zero time, the initial conditions are

$$t=0,\ I(0)=\frac{V_i}{R} \text{ and } \frac{d_i}{dt}=0$$

where $I(0)$ has the same meaning as ΔI in the paper.

A general solution for the current is

$$i=I_{\text{peak}}\epsilon^{-\alpha t}\sin(\omega t+\gamma)$$

and the voltage across the switch (across the diode) in this case is

$$V=\frac{1}{C}\int_0^t I_{\text{peak}}\epsilon^{-\alpha t}\sin(\omega t+\gamma)dt$$

where α is the damping factor, $R/2L$, and $\gamma=\tan^{-1}\omega/\alpha$.

Since in the recovery circuit R is relatively low, the approximation can be made that

$$\gamma=90^\circ,\ \omega=\frac{1}{\sqrt{LC}} \text{ and } I(0)=I_{\text{peak}}\times\sin\gamma=I(0)$$

and the voltage across the diode is

$$V=I(0)\frac{1}{C}\int_0^t\epsilon^{-\alpha t}\cos\omega t$$

$$=\frac{I(0)\epsilon^{-\alpha t}(\omega\sin\omega t-\alpha\cos\omega t)}{C(\alpha^2+\omega^2)}$$

Since $\omega\gg\alpha$

$$V=\frac{I(0)\epsilon^{-\alpha t}\sin\omega t}{C\omega}=I(0)\sqrt{\frac{L}{C}}\sin\omega t$$

An Experimental Study on the Ampere-Turn Gain of Ordinary Saturable Reactors

WILLIAM A. GEYGER
FELLOW AIEE

Synopsis: Actual deviations of the ordinary saturable-reactor control characteristic from the ideal 45-degree line have been measured directly by means of a differential-type test circuit, using bifilar saturable-reactor windings, which secure correct turns ratio and eliminate magnetic leakage. Three magnetic-amplifier core materials (Silectron, Orthonol, and Supermalloy) have been investigated in this way, with various values of the effective turns ratio between d-c control windings and a-c load windings.

The results of these measurements make it evident that quite remarkable performance of d-c instrument transformers may be obtained by slightly increasing the effective ratio of primary to secondary turns and by applying a properly rated a-c bias circuit, which shifts the control characteristic in such an amount that the error-current values are minimized in the entire working range of the saturable reactor.

Discussion of the results of these measurements is also concerned with performance of the respective saturable reactor utilizing external feedback or self-saturation. In this case, the actually measured deviations of the ordinary saturable-reactor control characteristic from the ideal 45-degree line (the "feedback line") become the very foundation on which such magnetic-amplifier circuits operate. These considerations include the typical S-shaped control characteristics causing bistable operation ("snap action"). By applying the novel test method described in this paper it is possible to measure, with a high degree of accuracy, the entire S-shaped path, along which the operating point of a bistable magnetic amplifier does actually move, when snap action occurs.

THE ampere-turn gain[1] is a fundamental quantity of ordinary saturable reactors. In the case of an ideal core material with a step-shape **B-H** characteristic (rectangular **B-H** line), the relation between load-winding ampere-turns, $I_L N_L$, and control-winding ampere-turns, $I_C N_C$, becomes a straight line, which for identical scales of ampere-turns subtends an angle of 45 degrees with the axes. Thus, in the idealized case, the ampere-turn gain represented by the slope of the control characteristic, $I_L N_L = f(I_C N_C)$, is $K_{AT} = (I_L N_L)/(I_C N_C) = 1$.

Performance of saturable reactors is governed by the deviations of the actually measured control characteristic from the ideal 45-degree line. These deviations are due to deviations of the actual dynamic **B-H** loop of the core material from the ideal, rectangular **B-H** line. Therefore, the closer the approximation between actual and idealized **B-H** characteristic, the higher will be the performance of the saturable reactor.

This fact is particularly significant with regard to "d-c instrument transformers,"[2-4] which are based upon the inherent current-transformer properties of ordinary saturable reactors with rectangular hysteresis-loop core material. Such arrangements make it possible to measure large direct currents without the use of shunt resistors; they permit measurement of high direct voltages in such a way that the meter circuit is isolated from the high-voltage circuit in accordance with conventional a-c instrument-transformer techniques. In these applications, it is the object of the designer to reduce the imperfections of the saturable reactor to the minimum and to approach the ideal operating conditions (the 45-degree line) as closely as possible.

The basic requirement of a very close approximation between the actual measured control characteristic of the ordinary saturable reactor and the 45-degree line is also of paramount importance with regard to magnetic amplifiers utilizing external-feedback windings, N_F, or self-saturation.[5] In the relevant feedback diagram[6,7] which characterizes the principle of operation of the self-saturating circuits and of equivalent external-feedback circuits with the feedback factor $h_s = N_F/N_L = 1$, the "feedback line," $I_L N_L = f(I_L N_F)$, is identical with the ideal 45-degree line of the respective ordinary saturable-reactor circuit. In this case, the deviations of the actually measured control characteristic, $I_L N_L = f(I_C N_C)$, of the ordinary saturable reactor from the theoretical 45-degree line are the basis of a simple transfer rule[8] for the graphical determination of the entire control characteristic of an external-feedback circuit with $N_F = N_L$.

It is to be noted that the closeness in which the characteristic $I_L N_L = f(I_L N_F)$ of an external-feedback circuit with $N_F = N_L$ approximates the theoretical 45-degree line is of decisive importance as regarding actual performance of "self-balancing" magnetic amplifiers[9-11] operating with infinite internal gain, "critical regeneration." As a result of the nonideal magnetic properties of core materials, the slope of this characteristic is always smaller than that of the theoretical 45-degree line. Thus, in practice it is necessary to increase the effective value of N_F correspondingly in order to reduce the control-winding current to the minimum.

In view of these considerations, the present experimental study on the ampere-turn gain of ordinary saturable reactors has been made with three different core materials: Silectron, Orthonol, and Supermalloy. Using the turns ratio $N_C/N_L = 1$, and other well-defined turns ratios (up to 1.05), actual difference between the average values of control current I_C and load current I_L was measured directly, and the beneficial effects of a properly rated a-c bias circuit was also investigated.

Comparison Between Actual Control Characteristic and Ideal Ampere-Turn Characteristic of Ordinary Saturable-Reactor Circuits

The object of the present study is to measure with a high degree of accuracy the deviations of the actual control char-

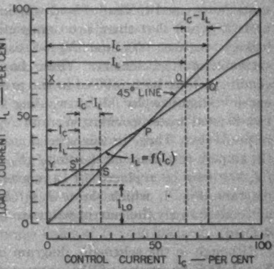


Fig. 1. Schematic diagram defining the deviations, $I_C - I_L$, of the actual control characteristic $I_L = f(I_C)$ from the 45-degree line

---

Paper 61-35, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted September 27, 1960; made available for printing November 18, 1960.

WILLIAM A. GEYGER is with the U. S. Naval Ordnance Laboratory, Silver Spring, Md.

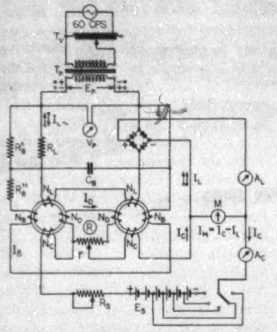


Fig. 2. Differential-type test circuit in which the zero-center-scale moving-coil instrument M indicates directly the error current, $I_M=I_C-I_L$, to be measured as a function of load current I_L

acteristic of ordinary saturable reactors from the ideal ampere-turn characteristic (the 45-degree line). Before describing the special test circuit used in these investigations, it is necessary to give a definition of the quantities to be measured. In the schematic diagram of Fig. 1, which is particularly suitable for this purpose, the deviations of the actual characteristic, $I_L=f(I_C)$, from the theoretical characteristic are greatly exaggerated to make a comparison of these characteristics and discussion of the essential points possible. Other graphical methods will be used in the following sections of this paper for presentation of the results of measurements, with actual scales corresponding to current ranges of instruments employed in the test circuit.

Definition of Current Values

Consider now the special case where the number of turns of control-winding units N_C is equal to that of load-winding units N_L, and that there is no magnetic leakage between N_C and N_L of each saturable-reactor element. These basic conditions have been fulfilled in the present investigations by bifilar winding of two twisted wires providing N_C and N_L, respectively. Thus, in this special case, the ampere-turn diagram showing $I_LN_L=f(I_CN_C)$ can be replaced by a current diagram, Fig. 1, which shows $I_L=f(I_C)$ and contains only the current values to be defined.

Although in the schematic diagram of Fig. 1 the deviations of the characteristic $I_L=f(I_C)$ from the 45-degree line are greatly exaggerated with regard to difference in slope and magnitude of the exciting current (I_{LO}) of the saturable-reactor elements, this presentation indicates quite correctly the general shape and position of the actual characteristic $I_L=f(I_C)$ with respect to the 45-degree line. Moreover, this diagram has been designed merely as an aid for describing qualitatively the relations of those current values, which are the basis of the following discussions.

The intersection, P, of the characteristic $I_L=f(I_C)$ with the 45-degree line signifies a reference point, where $I_C=I_L$. That is, at this point, the saturable reactor has ideal operating conditions. In the region above P, for example at the points Q and Q', the required control current $I_C=XQ'$ is larger than the associated ideal value $I_L=XQ$, and the deviation $I_C-I_L=QQ'$ is positive. However, in the region below P, for example at the points S and S', the required control current $I_C=YS'$ is smaller than the associated ideal value $I_L=YS$, and the deviation $I_C-I_L=SS'$ is negative.

Referring to the application of the ordinary saturable reactor as a d-c instrument transformer, the deviation I_C-I_L represents an error current, and the percentage current-ratio error in terms of the control current I_C to be measured, i.e., the deviation of the actual current ratio $K_I=I_L/I_C$ from its ideal value, $I_L/I_C=N_C/N_L=1$, is

$$\Delta K_I=100\frac{I_L-I_C}{I_C} \qquad (1)$$

This error ΔK_I is positive when, with a given "primary" current I_C, the actual "secondary" current I_L is in excess of its ideal value, i.e., when the ammeter indicating I_L reads too high.

If self-saturation or external feedback with $N_F=N_L$ is used, i.e., if the extrinsic positive feedback ($h_e=+1$) cancels completely the intrinsic negative feedback ($h_i=-1$) of the ordinary saturable reactor ($h=h_e+h_i=0$), then the deviations I_C-I_L, as defined in Fig. 1, become the very foundation on which these arrangements operate. In this case, the intersection P of the 45-degree line (the "feedback line") with the control characteristic $I_L=f(I_C)$ signifies a point where load current is obtained with zero average control current. This point is here termed the "quiescent point."

Since the theoretical ampere-turn gain of an ordinary saturable reactor with ideal core material is $K_{AT}=1$, in accordance with the ideal 45-degree line, the actual ampere-turn gain $K_{AT,SR}$ which is derived from the actually measured control characteristic, can be expressed by

$$K_{AT,SR}=1-\delta \qquad (2)$$

where δ designates the deviation of the slope of this characteristic from the 45-degree line. It may be stressed here, again, that in the schematic diagram of Fig. 1 the deviations I_C-I_L, corresponding to $\delta=1-K_{AT,SR}$ are greatly exaggerated to facilitate discussion of the essential points. Actually, when using nickel-iron-alloy core materials and proper core construction, the deviation δ in the linear part of the characteristic will be in the range of 0.002 to 0.005.

In view of this basic concept,[1] actual ampere-turn gain K_{AT} of the external-feedback circuit with $h_e=N_F/N_L$ and ideal rectifier properties (zero-reverse current) is given by the relationship

$$K_{AT}=K_{AT,SR}\frac{1}{1-h_eK_{AT,SR}}$$

$$=K_{AT,SR}\frac{1}{1-(1-\delta)N_F/N_L} \qquad (3)$$

With compensated feedback[6] ($h_e=N_F/N_L=1$) and in self-saturating circuits, where h_e is inherently equal to unity, the simplified expression

$$K_{AT}=K_{AT,SR}\frac{1}{1-K_{AT,SR}}=\frac{1-\delta}{\delta}\approx\frac{1}{\delta} \qquad (4)$$

may be used.

Thus, when measuring the deviations I_C-I_L (Fig. 1), corresponding to $\delta=1-K_{AT,SR}$ (equation 2), as a function of load current I_L, with a sufficiently high degree of accuracy, it will be possible to determine the actual control characteristic of a self-saturating circuit or of the equivalent external-feedback circuit with $N_F=N_L$. In this way, different core materials and core configurations may be studied and compared with regard to these deviations, which are a measure for the deviation of the actual dynamic **B-H** loop of the core from the ideal rectangular **B-H** line.

As regarding "self-balancing" magnetic amplifiers,[9–11] the deviations I_C-I_L represent an error current; and $(I_C-I_L)=f(I_L)$ characterizes, in this case, actual deviations of the control-winding current from its ideal value, $I_C-I_L=0$. When using other well-defined turns ratios N_C/N_L (e.g., larger values of N_C/N_L up to 1.05, including "fractional turns,") the effects of additional positive feedback can be investigated, particularly for "infinite gain." Comparison of different core materials and core configurations with respect to actual values of the deviations I_C-I_L as a function of load current I_L makes it possible, in this case, to determine the error-current characteristic, which is a criterion for the performance of the cores in a self-balancing magnetic-amplifier circuit.

In view of the beneficial effects of a properly rated a-c bias circuit measurement of the error-current characteristic $I_C - I_L = f(I_L)$ is particularly important. Application of a-c bias has proved to be definitely advantageous for d-c instrument transformers and low-level self-balancing magnetic amplifiers; it may be quite valuable in many other applications of saturable-reactor arrangements. For this reason, operation of the ordinary saturable reactor with a-c bias has been included in the present investigations.

THE TEST METHOD

The specific instrumentation problem consists in measuring with a sufficiently high degree of accuracy the actual difference of the average values of currents I_C and I_L, which may be either positive or negative. Obviously, measurement of the current difference $I_C - I_L$ by means of two directly indicating moving-coil-type ammeters, even when the best grade instruments may be used, involves practical difficulties because the actual changes in pointer deflection of such instruments, only about 0.1 to 2%, are very small. In order to avoid these difficulties, a special differential-type method, similar to well-known test methods for a-c instrument transformers, must be used.

A first step in this direction was the application of a d-c potentiometer method with high-precision standard resistors. This null method[12] makes it possible to measure current ratio I_L/I_C and ampere-turn ratio $(I_L N_L)/(I_C N_C)$ with an absolute accuracy of ±0.02 to 0.03%.

Another suitable method which has been used for testing of a miniaturized d-c instrument transformer,[13] with the turns ratio $N_C/N_L = 10{,}000/10{,}000 = 1$, consists in superposing load current I_L and control current I_C in a zero-center-scale moving-coil instrument which indicates directly the difference current $I_C - I_L$, which is a measure for the percentage current-ratio error in terms of I_C, as defined by equation 1. Imperfections of this instrument enter into the measurement as errors of second order. Therefore, it is possible to employ an ordinary panel-type microammeter for measuring ΔK_I with an accuracy of about ±1 to 2%.

In the present study, a similar deflection method was used for measuring the difference current $I_C - I_L$ as a function of I_L. As mentioned before, equality of the actually applied number of turns of winding units, N_C and N_L, and the elimination of magnetic leakage between these winding units has been secured in an ideal way by bifilar winding of two twisted wires providing N_C and N_L, respectively.

Fig. 2 shows schematically the test circuit with the two equally rated tape-core elements having series-opposing-connected a-c windings (load windings N_L, bias windings N_B) and series-aiding-connected d-c windings (main control windings N_C, additional control windings N_D). All these windings consist of six twisted wires, which are hand-wound (e.g., with 500 turns on each core element) and distributed as follows:

Load windings N_L, 500+500=1,000 turns
Bias windings N_B, 500 turns
Main control windings N_C, 500+500=1,000 turns
Additional control windings N_D, 500 turns

The a-c load resistor R_L and a bridge-type full-wave rectifier consisting of four silicon-junction diodes are series-connected with N_L. Thus, the unidirectional load current I_L flowing through the moving-coil ammeter A_L represents the average value of the full-wave-rectified alternating current, $I_L\sim$, which flows through N_L. Actual average value of the 60-cycle power-supply voltage E_P, indicated by a rectifier-type voltmeter V_P, can be varied by means of a continuously adjustable transformer T_V feeding the isolation transformer T_P.

Load current I_L and control current I_C are superimposed in a moving-coil zero-center-scale milliammeter or microammeter M, which directly indicates the error current $I_M = I_C - I_L$ to be measured as a function of I_L. In order to ensure operation of the test circuit with free even-harmonic currents ("natural" magnetization conditions), the number of active cells of the storage battery, i.e., actual magnitude of control or signal voltage, E_S, is always selected so that the variable series resistance R_S has a very small value. As a result, the superimposed currents I_L and I_C have substantially the same waveform, and the pointer of M does not show excessive vibrations (such vibrations may also be eliminated by means of a large capacitor connected across the terminals of M).

Although the reference turns ratio, $N_C/N_L = 1$, is generally used for measuring the deviations of the control characteristics $I_L = f(I_C)$ from the 45-degree line, other well-defined turns ratios (up to 1.05, including "fractional turns") may be provided to study correspondingly modified operating conditions of the saturable reactor. This is made possible by employing additional control-winding units N_D in conjunction with an Ayrton-Mather shunt. In accordance with Fig. 2, the effective control ampere-turns are given by the expression

$$I_C N_{C,\mathrm{EFF}} = I_C N_C + I_D N_D = I_C \left(N_C + N_D \frac{r}{R} \right) \tag{5}$$

where R represents the total loop resistance of the current-divider circuit, and r is the active part of the potentiometer resistor, as determined by the actual position of the slider. Alternately, turns ratios $N_{C,\mathrm{EFF}}/N_L < 1$ may be provided by reversing the terminals of N_D:

$$I_C N_{C,\mathrm{EFF}} = I_C N_C - I_D N_D = I_C \left(N_C - N_D \frac{r}{R} \right) \tag{6}$$

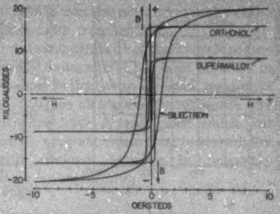


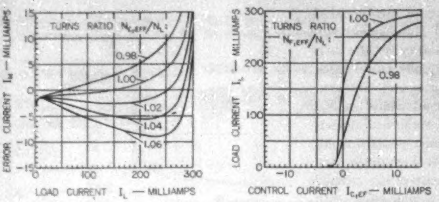
Fig. 3. 60-cps B-H loops of the three core elements which have been investigated in the test circuit of Fig. 2

The a-c bias circuit[14] consists of bias windings N_B, series resistors R_B', R_B'', and shunt capacitor C_B. It is to be noted that, with low-coercive-force nickel-iron-alloy core materials (e.g., Supermalloy or Mumetal) it is possible to obtain the desired reduction of the exciting current, I_{LO} in Fig. 1, only if the phase relationship between bias current I_B and power-supply voltage E_P is properly adjusted by varying the values of R_B', R_B'', and C_B. When testing grain-oriented core materials (e.g., Orthonol or Deltamax), the capacitor C_B may be generally omitted. However, with Silectron cores, application of C_B is necessary to obtain optimum value of phase displacement between the fundamental waves of current I_B and voltage E_P.

The fundamental requirement of perfect full-wave rectification of the alternating load current $I_L\sim$ is met by employing silicon-junction diodes having a low forward resistance and an extremely high reverse resistance. Consequently, the assumption that I_L represents the average value of $I_L\sim$ is justified.

Discussion of Experimental Results

Three different magnetic-amplifier core materials (Silectron, Orthonol, and Super-

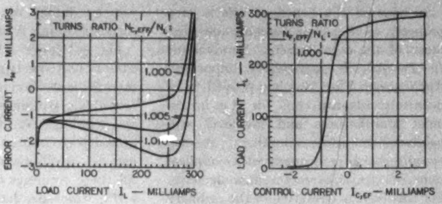


(A) (B)

Fig. 4. Results of measurements with Silectron cores, without a-c bias

A—Error-current characteristic, $I_M=f(I_L)$ of d-c instrument transformer with turns ratio $N_{C,EFF}/N_L$ as a parameter

B—Corresponding control characteristic, $I_L=f(I_{C,EF})$ of an external-feedback circuit with feedback turns ratio, $N_{F,EFF}/N_L$ as a parameter

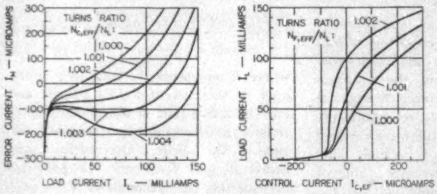


(A) (B)

Fig. 5. Results of measurements with Orthonol cores, without a-c bias

A—Error-current characteristic, $I_M=f(I_L)$ of d-c instrument transformer with turns ratio $N_{L,EFF}/N_L$ as a parameter

B—Corresponding control characteristic $I_L=f(I_{C,EF})$ of an external-feedback circuit with feedback turns ratio $N_{F,EFF}/N_L=1$

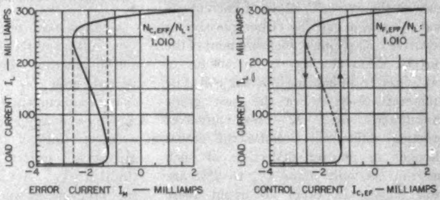


(A) (B)

Fig. 6. Results of measurements with Supermalloy cores, without a-c bias

A—Error-current characteristic, $I_M=f(I_L)$ of d-c instrument transformer with turns ratio $N_{C,EFF}/N_L$ as a parameter

B—Corresponding control characteristic, $I_L=f(I_{C,EF})$ of an external-feedback circuit with feedback turns ratio $N_{F,EFF}/N_L$ as a parameter



(A) (B)

Fig. 7. S-shaped characteristic of the saturable-reactor circuit with Orthonol cores

A—Actually measured error-current characteristic, $I_L=f(I_M)$, of the ordinary saturable-reactor circuit with turns ratio $N_{C,EFF}/N_L=1.010$

B—Derived looped control characteristic, $I_L=f(I_{C,EF})$, of the corresponding bistable external-feedback circuit with turns ratio $N_{F,EFF}/N_L=1.010$

malloy) have been investigated using the test circuit of Fig. 2. The following discussion of experimental results presents the actual performance characteristics of these materials in an ordinary series-connected saturable reactor acting as a d-c instrument transformer. It also permits comparison of these materials with respect to their performance in external-feedback circuits or self-saturating circuits (control characteristic) and in self-balancing magnetic-amplifier circuits (error-current characteristic).

DIMENSIONS OF COMPONENTS

The Orthonol 2-mil-tape cores and the Supermalloy 2-mil-tape cores have the same dimensions: inside diameter = 1.25 inches, outside diameter = 1.75 inches, tape width = 0.50 inch. Each core has 6×500 turns of no. *30* (American Wire Gage) wire (6 twisted wires, hand-wound, 6×15 ohms). These windings are connected so that $N_L=1{,}000$ turns, $N_C=1{,}000$ turns, $N_D=500$ turns, and $N_B=500$ turns, as already described in the preceding section. Thus the measurements with Orthonol and Supermalloy core elements have been made under almost identical conditions with regard to the dimensions of the cores and of the windings; only the core material has been changed to get the net differences between actual performance characteristics using the differential-type test circuit of Fig. 2.

The dimensions of the Silectron 12-mil-tape core elements (type *T 10241*,* *L12V*, uncut) are slightly larger than those of the Orthonol and Supermalloy core elements. However, actual differences in cross-sectional area and mean length of the magnetic path are small enough to permit this general comparison of performance characteristics with the power-supply frequency $f_P=60$ cps (cycles per second). Fig. 3 presents the 60-cps hysteresis loops of these core elements, which have been automatically ink-recorded by means of a simple dynamic **B-H** loop tracer with silicon-diode chopper circuits (Ferrotracer)[16–17] and a standard chopper-type *x-y* plotter. The magnetic-core test circuit itself has been designed so that the core elements are tested under sinusoidal-flux excitation conditions.

The full-wave rectifier bridge circuit consists of four silicon-junction diodes (e.g., type *1 N 458*). Results of measurements with four germanium diodes (type *GE 1 N 93**) have shown that these diodes may be used equally well in the test circuit of Fig. 2. The maximum value of load current I_L was 150 milliamperes, or 300 milliamperes, respectively. The total load-circuit resistance was determined in each case, and the power-supply voltage E_P was adjusted so that the alternating fluxes in the core elements have a peak value which just fails to saturate either core, if the control current I_C is zero.

RESULTS OF MEASUREMENTS

Fig. 4(A) shows for the Silectron cores error current $I_M=I_L-I_L$ as a function of I_L with turns ratio $N_{C,EFF}/N_L$ (equa-

---

* Arnold Engineering Company.

* General Tire Company.

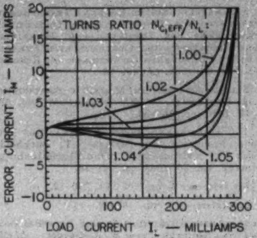


Fig. 8. Results of measurements with Silectron cores using a-c bias circuit: Error-current characteristic $I_M=f(I_L)$ of d-c transformer with turns ratios $N_{C,EFF}/N_L=1.00$ to 1.05

tions 5 and 6) as a parameter. The modified presentation of Fig. 4(B) gives in the usual form the control characteristic of an external-feedback circuit, i.e., load current I_L as a function of the respective control current, $I_{C,EF}$ with the corresponding turns ratio $N_{F,EFF}/N_L$ as a parameter. Figs. 5 and 6 show the respective characteristics for Orthonol and Supermalloy cores.

It is to be noted that Figs. 4(B), 5(B), and 6(B) show only those control characteristics which represent "stable" operation of the respective external-feedback circuit (without snap action). Obviously, the S-shaped characteristics of Figs. 4(A), 5(A), and 6(A) will cause "bistable" operation of the external-feedback circuit (with snap action).

Fig. 7(A) shows as an example, for the Orthonol cores with the effective turns ratio $N_{C,EFF}/N_L=1.010$, the S-shaped ordinary-saturable-reactor control characteristic which has been measured with a high degree of accuracy. The dashed vertical lines have been added here to indicate those control-current values at which snap action will occur in the respective external-feedback circuit. In the modified diagram of Fig. 7(B) with the corresponding turns ratio $N_{F,EFF}/N_L=1.010$, the vertical lines (with arrows showing the direction of snap action) are parts of the looped control characteristic $I_L=f(I_{C,EF})$ resulting from bistable operation of the external-feedback circuit. Because stable operating conditions between the upper and lower values of load current I_L are not possible, the S-shaped path along which the operating point does actually move, when snap action occurs, is indicated in Fig. 7(B) by dashes.

The results of measurements presented in Figs. 4(A), 5(A), and 6(A) make it evident that quite remarkable performance of the saturable reactor acting as a d-c instrument transformer may be obtained by slightly increasing the effective ratio $N_{C,EFF}/N_L$ of "primary" to "secondary" turns, in accordance with common practice on a-c instrument transformers. It is also evident from the d-c instrument-transformer characteristics of Figs. 4(A), 5(A), and 6(A) that performance may be considerably improved by shifting the control characteristic in such an amount that the error-current values (I_M) are minimized in the entire working range of the circuit. Figs. 8 through 10 show corresponding performance characteristics, $I_M=f(I_L)$, for the case that a proper amount of a-c bias is applied to obtain minimized values of I_M. Fig. 8 makes it evident that the d-c instrument transformer with Silectron cores behaves remarkably well, when a properly rated a-c bias circuit and an optimized value of turns ratio $N_{C,EFF}/N_L$ are used.

Influence of Changes in Load-Circuit Resistance and Variations in Power-Supply Voltage

Optimized performance (minimum values of I_M) can be obtained by reducing the total resistance of the load circuit to the minimum and by applying a substantially constant power-supply voltage (E_P), preferably to be derived from a ferroresonant-type constant-voltage transformer. In this way, it is possible to take full advantage of optimized operating conditions.

It is to be noted that, in the present study, the measurements with the test circuit of Fig. 2 have been made deliberately employing a comparatively large value of total load-circuit resistance (about 100 ohms) to demonstrate the fact of increasing values of error current I_M with large values of load current I_L. In practice, however, much smaller values of load-circuit resistance may be used to obtain even higher performance of d-c instrument transformers.

Actual magnitude of the 60-cycle power-supply voltage (E_P) has been adjusted in all cases so that, with no-signal conditions ($I_C=0$), the alternating fluxes in the core elements have a peak value which just fails to saturate either core, in accordance with common practice in magnetic-amplifier engineering.

Summary and Conclusions

Actual deviations of the ordinary-saturable-reactor control characteristic from the ideal 45-degree line have been measured directly by means of a differential-type test circuit, using bifilar saturable-reactor windings, which secure correct turns ratio and eliminate magnetic leakage. Three magnetic-amplifier core materials (Silectron, Orthonol, and Supermalloy) have been investigated in this way, with various values of the effective turns ratio between d-c control windings and a-c load windings. The results of measurements may be summarized as follows:

1. Without using a-c bias, quite remarkable performance characteristics of the ordinary saturable reactor acting as a d-c instrument transformer may be obtained by slightly increasing the effective ratio ($N_{C,EFF}/N_L$) of "primary" to "secondary" turns, in accordance with common practice on a-c instrument transformers.
2. The error-current characteristic of d-c instrument transformers may be considerably improved by shifting the entire characteristic to such an amount that the error current ($I_M=I_C-I_L$) is minimized in the entire working range of the saturable reactor. This can be achieved easily by applying a properly rated a-c bias circuit.

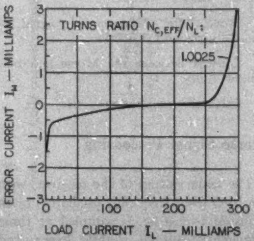


Fig. 9. Results of measurements with Orthonol cores using a-c bias circuit: Error-current characteristic $I_M=f(I_L)$ of d-c transformer with turns ratio $N_{C,EFF}/N_L=1.0025$

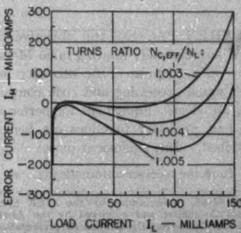


Fig. 10. Results of measurements with Supermalloy cores using a-c bias circuit: Error-current characteristic $I_M=f(I_L)$ of d-c transformer with turns ratios $N_{C,EFF}/N_L=1.003$ to 1.005

3. Optimized performance (minimum values of I_M) can be obtained by reducing the total resistance of the load circuit to the minimum and by applying a substantially constant power-supply voltage (E_P), preferably to be derived from a ferroresonant-type constant-voltage transformer. In this way, it is possible to take full advantage of optimized operating conditions.

4. Another important requirement for minimizing the error current, I_C-I_L, consists in adjusting actual magnitude of power-supply voltage (E_P) so that, with no-signal conditions ($I_C=0$), the alternating fluxes in the core elements have a peak value which just fails to saturate either core, in accordance with common practice in magnetic-amplifier engineering.

5. With Silectron cores, effective turns ratio $N_{C.EFF}/N_L=1.05$, and optimum adjustment of the a-c bias circuit, Fig. 8, actual error current $I_M=I_C-I_L$ varies within the limits of ±0.8% of the assumed maximum value of load current I_L (250 milliamperes) which corresponds to full-scale deflection of the ammeter indicating I_L.

6. The corresponding error-current value for Orthonol cores and $N_{C.EFF}/N_L=1.0025$, Fig. 9, varies within the limits of ±0.4% of the maximum value of I_L (250 milliamperes).

7. The corresponding error-current value for Supermalloy cores and $N_{C.EFF}/N_L=1.004$, Fig. 10, varies within the limits of ±0.1% of the maximum value of I_L (150 milliamperes).

Referring to the modified presentation of experimental results, Figs. 4(B), 5(B), and 6(B), in the form of a control characteristic, $I_L=f(I_{C.EF})$ of an external-feedback circuit with the turns ratio $N_{F.EFF}/N_L$, it is evident that the present test method provides an instructive way for demonstrating the extreme significance of actual deviations of the control characteristic $I_L=f(I_C)$ from the 45-degree line with regard to the fundamental principle of external-feedback-type and self-saturating magnetic-amplifier circuits. These considerations include the typical S-shaped control characteristic causing bistable operation (snap action). By applying the novel test method described in this paper it is possible to measure, with a high degree of accuracy, the entire S-shaped path, along which the operating point of a bistable magnetic amplifier does actually move, when snap action occurs.

References

1. Magnetic Amplifiers (book), H. F. Storm. John Wiley & Sons, Inc., New York, N. Y., 1955, pp. 107–08, 206, 210.

2. A Simple D-C Instrument Transformer Having Real Current-Transformer Properties (in German), W. Krämer. *Elektrotechnische Zeitschrift*, Wuppertal-Elberfeld, Germany, vol. 58, 1937, pp. 1309–13.

3. A New D-C Instrument Transformer for the Measurement of High D-C Voltages (in German), W. Krämer. *Ibid.*, vol. 59, 1938, pp. 1295–98.

4. Separately Excited Current Transformer Operating as a Universal-Type Instrument Transformer for Oscillographic Investigations on Alternating Currents with Direct-Current Components (in German), W. Krämer. *Ibid.*, vol. 60, 1939, pp. 393–95.

5. Self-Saturation in Magnetic Amplifiers, W. J. Dornhoefer. *AIEE Transactions*, vol. 68, pt. II, 1949, pp. 835–50.

6. D-C Controlled Saturable Reactors and Their Feedback (in German), T. Buchhold. *Archiv für Elektrotechnik*, Berlin-Charlottenburg, Germany, vol. 36, 1942, pp. 221–38.

7. See reference 1, pp. 197–201, 210–12.

8. See reference 1, pp. 197–99.

9. Magnetic Amplifiers of the Self-Balancing Potentiometer Type, W. A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 71, 1952 (Jan. 1953 section), pp. 383–95.

10. Self-Balancing Magnetic Amplifier, W. A. Geyger. *U. S. Patent no. 2,700,130*, Jan. 18, 1955.

11. The Effective Feedback Factor of Self-Balancing Magnetic Amplifiers, W. A. Geyger. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Mar. 1958, pp. 66–69.

12. Magnetic Amplifiers with Orthonol Tape Cores, W. A. Geyger. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 59–67.

13. A Miniaturized D-C Instrument Transformer, William A. Geyger. "1959 Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication T116*, Sept. 1959, pp. 130–53.

14. Magnetic-Amplifier Circuits (book), W. A. Geyger. McGraw-Hill Book Company, Inc., New York, N. Y., second edition, 1957, pp. 78–81, 118–19, 123–25.

15. The Ferrotracer, A Simple Dynamic B-H Loop Tracer with Silicon-Diode Chopper Circuits, W. A. Geyger. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 12, 1956, pp. 83–97.

16. Magnetic-Switch B-H Loop Tracer, W. A. Geyger. *Electronics*, New York, N. Y., vol. 29, no. 10, Oct. 1956, pp. 167–69.

17. See reference 14, pp. 350–69, fig. 20.10.

Transistor Morse-to-Teleprinter Code Converter

J. F. CUNNIFF
NONMEMBER AIEE

C. E. THEALL, JR.
NONMEMBER AIEE

WHEN FACING the objective of automatically printing radio Morse copy on a teleprinter, the necessity for both signal processing and code conversion are clear. Equipment for performing four necessary functions are herein described. This equipment must:

1. Tune the receiver automatically.
2. Process the Morse-coded signal to remove noise and distortion.
3. Convert between the Morse and teleprinter codes.
4. Drive the teleprinter.

Morse Signal Processing

The audio output of the receiver, with the BFO (beat frequency oscillator) pitch adjusted to 1.0 kc, is sampled at three frequencies: 1.0 kc, slightly lower, and slightly higher. A mark is judged to be present when the detected output of the center channel exceeds the average of the side channels by an arbitrary amount. Noise bursts generally produce equal outputs in all three channels, and this appears as a shift in the base line. The signal plus noise of the center channel output is still distinguishable from the noise outputs of the side channels when a mark is present. Switching of the detector characteristics with varying reception conditions provides additional immunity to the effects of interfering signals. For those sets of signal conditions in which the channel filter circuits deliver an occasional mark which is excessively shorter than the dot length for the message rate being received, accessory circuitry discriminates against the short mark.

Automatic Receiver Tuning

Automatic tuning of the receiver is accomplished with a small servomotor assembly mounted on the receiver in place of the BFO tuning knob. The receiver can be mounted remote from the code converter. The control signal for the motor

---

Paper 61-98, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted March 9, 1960; made available for printing January 10, 1961.

J. F. Cunniff is with Trak Electronics Company, Inc., Wilton, Conn. C. E. Theall, Jr., formerly with Trak Electronics Company, is with Wertheim and Company, New York, N. Y.

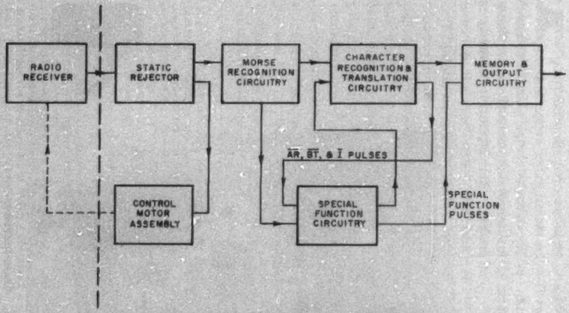

RADIO RECEIVER
STATIC REJECTOR
CONTROL MOTOR ASSEMBLY
MORSE RECOGNITION CIRCUITRY
CHARACTER RECOGNITION & TRANSLATION CIRCUITRY
$\overline{AR}$, $\overline{BT}$, & $\overline{I}$ PULSES
SPECIAL FUNCTION CIRCUITRY
SPECIAL FUNCTION PULSES
MEMORY & OUTPUT CIRCUITRY

Fig. 1. Code conversion system block diagram

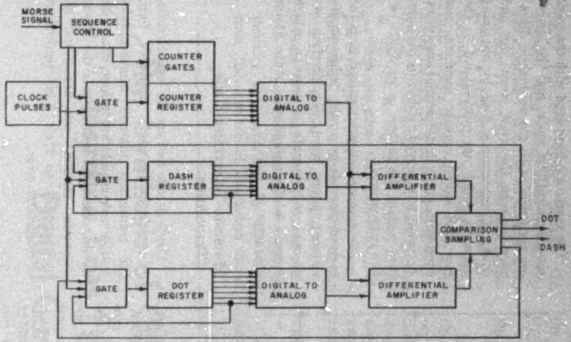

MORSE SIGNAL
SEQUENCE CONTROL
CLOCK PULSES
GATE
COUNTER GATES
COUNTER REGISTER
DIGITAL TO ANALOG
GATE
DASH REGISTER
DIGITAL TO ANALOG
DIFFERENTIAL AMPLIFIER
COMPARISON SAMPLING
DOT
DASH
GATE
DOT REGISTER
DIGITAL TO ANALOG
DIFFERENTIAL AMPLIFIER

Fig. 3. Morse recognition for dots and dashes

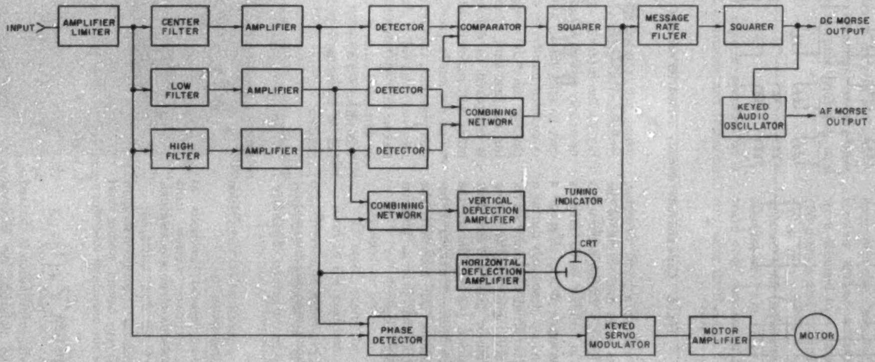

INPUT
AMPLIFIER LIMITER
CENTER FILTER
LOW FILTER
HIGH FILTER
AMPLIFIER
AMPLIFIER
AMPLIFIER
DETECTOR
DETECTOR
DETECTOR
COMPARATOR
COMBINING NETWORK
COMBINING NETWORK
VERTICAL DEFLECTION AMPLIFIER
HORIZONTAL DEFLECTION AMPLIFIER
TUNING INDICATOR
CRT
SQUARER
MESSAGE RATE FILTER
SQUARER
DC MORSE OUTPUT
KEYED AUDIO OSCILLATOR
AF MORSE OUTPUT
PHASE DETECTOR
KEYED SERVO MODULATOR
MOTOR AMPLIFIER
MOTOR

Fig. 2. Static rejector block diagram

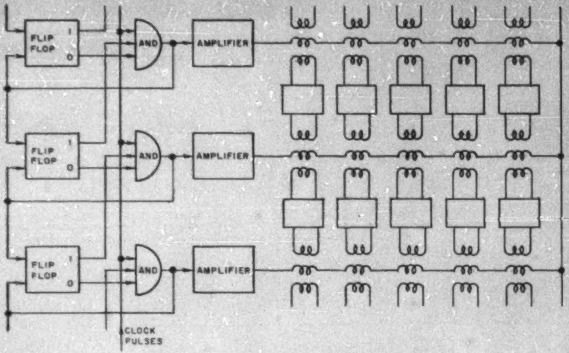


Fig. 4. Code converter memory 3-character columns

servo is derived from a phase detector, operating across input and output of the center channel filter. Tuning accuracy is ±3 cycles. The tuning servo loop is closed only during a mark. Use of a motor servo in a loop closed only during a mark, rather than an all-electronic servo, provides long-term memory in the tuning process. The motor shaft will not drift off to an unknown position during the interval between Morse characters or between transmissions, but maintains the receiver tuning at the optimum position for the last mark received.

Translator

The translator converts Morse code into teleprinter code at average word-per-minute rates up to the printer capability. Features are:

1. Automatic regulation of recognition speed to input code rates.
2. Generation of special functions as needed to form teleprinter controls: LINE FEED, CARRIAGE RETURN, FIGURES, LETTERS, and SPACE.
3. Teleprinter character memory to allow the translator to copy behind, storing short characters while the teleprinter catches up.

Morse Recognition

In Morse code, information is derived from the signal polarity, the interval duration, and the number of pulses. Recognition is the process of measuring and identifying dots, dashes, and element, letter, and word spaces. The marks are dots and dashes; 1 and 3 units long respectively. The spaces are element, letter, and word spaces; 1, 3, and 7 units long respectively.

A Morse interval gates a train of pulses, which are counted in a counter-register circuit and are given a number corresponding to the interval; gating circuits momentarily switch the counter-register circuit to shift mode, and the recently accumulated number is shifted end around through the counter-register twice. During the first shift operation, the number is compared to numbers stored in two other shift registers, which contain a previous dot and a previous dash. During this first shift, the identity of the recently received interval is determined; in the second shift operation, an identifying pulse is generated and, if the latest interval was identified as a dot or dash, its number is transferred to the dot or dash register.

The marks are compared with the most recently identified mark, and the spaces are compared with the arithmetic difference between the most recent dot and dash. Comparison is done by obtaining voltages, which are proportional to the new and stored register numbers, and comparing them in differential amplifiers. Accuracy is improved by shifting the numbers toward the highest order digits before comparison.

The *comparisons* are sampled near the ends of the code intervals; pulses signify element, letter, word, dot, or dash decisions, and operate a register which stores a 7-bit representation of the Morse character.

Morse-to-Teleprinter Code Translation

Diode decode and encode logic circuits, connected by transistor amplifiers, form the translation matrix. Morse-to-teleprinter code translation is done, character by character; the seven bits of the Morse character register are translated simultaneously to their 5-bit equivalent in teleprinter code.

Teleprinter Character Memory

A magnetic core buffer memory couples the translation matrix to the teleprinter amplifier; ten teleprinter characters can be stored in an array, 5 cores wide and 10 columns long. The distinguishing feature of the memory is its character process. A teleprinter character nor-

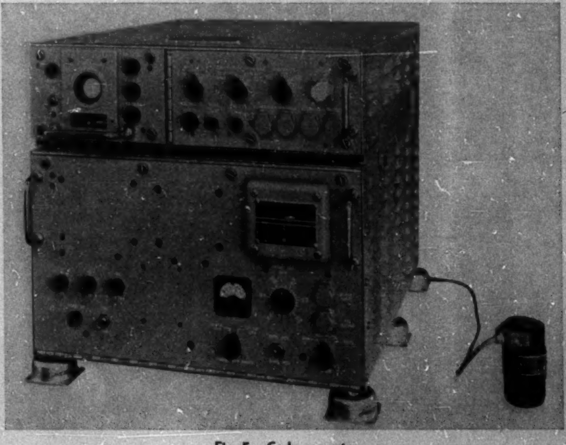


Fig. 5. Code converter

mally transfers rapidly through the memory, being shifted along by memory clock pulses. Near the output end of the memory or during bursts of short characters, a character may be stored in a column for a period appreciably longer than a clock pulse period. New characters cannot be transferred to a column during its character-stored state because of the column flip-flop's control of the previous AND gate—that gate whose output is energized only when every input is energized. Also, a stored character cannot transfer from the column if the next column is in the character-stored state. The rate of character flow through the memory is determined by the signal rate, the number of characters stored, the teleprinter speed, and, to some extent, the clock frequency.

Special Functions

Special function circuitry inserts additional characters into the magnetic memory stages. These special characters are not present in the Morse signal, but are derived from various criteria and generate the output characters: SPACE, LINE FEED, CARRIAGE RETURN, FIGURES, and LETTERS.

Output Circuit

A serial shift register operates as a parallel-to-serial converter and transfers the contents of the tenth memory column to the output amplifier. Reshaping of the register pulses by gates and multivibrators forms a squared signal, which keys a carrier-coupled teleprinter line amplifier.

Conclusions

The objectives of copying in the presence of noise, automatic tuning of the receiver, and a permanent message record were attained in a mobile unit weighing 120 pounds and of modular construction which provides good packing and low volume while retaining accessibility for efficient servicing. Figs. 1 through 5 illustrate code conversion processes and the converter itself

---

The Synthesis of General Parameter Insertion Loss Filters Using a Digital Computer

G. C. TEMES
NONMEMBER AIEE

A PROBLEM often encountered in practical network design is the synthesis of filters which satisfy various frequency-dependent requirements in the stop band and exhibit Tchebyshev behavior in the pass band. The theory of such networks, usually called general parameter filters, was founded by Darlington[1] and treated in detail by German authors.[2–4] As the amount of numerical calculations encountered in the design process is considerable even for the simplest cases and becomes excessive for the more involved networks, the need for a digital computer is strongly indicated. As will be pointed out later, a digital computer supplies the fastest and most economical way of design.

General Parameter Filters

Most of the requirements for general parameter filters can be classified under one of the following types:

1. Prescribed attenuation poles in stop band, prescribed pass-band ripple.
2. Given required minimum attenuation (as a function of frequency) in the stop band, prescribed pass-band ripple.

For case 2, the attenuation requirements in the stop band can be given in a graphical or verbal form [e.g., $A>40$ db (decibels) for $0<f<f_1$, $A>20$ db for $f_1<f<f_2$, $A>60$ db for $f_2<f<f_s$].

As case 2 can be reduced to case 1 by determining the suitable location of poles through graphical methods,[2,5,6] only the synthesis of filters satisfying type 1 requirements will be treated here. Filters of this class have considerable importance of their own, involving such uses as the suppression of unwanted products in frequency changers, the blocking of signaling and pilot frequencies, etc., while passing the band of useful signals undistorted.

The following is a short summary of the design of these filters restricted to symmetrical networks. The treatment of antimetric filters follows substantially similar lines.

The insertion power ratio for a filter of this type is given by

$$e^{2\alpha}=1+(e^{2\alpha_p}-1)\sinh^2\theta_I$$

where α is the insertion loss function and α_p is the pass-band ripple, both in nepers, and $\theta_I(p)$ is a design parameter. Following Darlington,[1] $\theta_I(p)$ is chosen as the image transfer function of an antimetric image parameter filter, which has a theoretical pass band coinciding with the specified pass band as well as the prescribed attenuation poles.

It could easily be proved that this form of the insertion power ratio leads generally to realizable networks, satisfying both the pass- and stop-band requirements.

Using the Q' functions[3] to express θ_I,

$$e^{2\alpha}=1+(e^{2\alpha_p}-1)\times \left|\frac{p[b_0+b_2p^2+b_4p^4+\ldots+b_{2\sigma}p^{2\sigma}]}{(1+z_1p^2)(1+z_2p^2)\ldots(1+z_\sigma p^2)}\right|^2$$

Here $p=j\Omega$ is the normalized complex frequency and the $b-s$ are constants, given by

$$b_0=1+2\delta_1$$

$$b_{2\mu}=\binom{\sigma}{\mu}+\sum_{\tau=1}^{\mu}\binom{\sigma-\tau}{\mu-\tau}\times \left[\delta_\tau^2+2\sum_{k=1}^{\tau \text{ or } \sigma-\tau^*}\delta_{\tau+k}\delta_{\tau-k}\right]+ 2\sum_{\tau=0}^{\mu}\binom{\sigma-\tau}{\mu-\tau}\sum_{k=1}^{\tau+1 \text{ or } \sigma-\tau^*}\delta_{\tau-k}\delta_{\tau-k+1}$$

where the asterisk (*) indicates that whichever is the smaller should be used, and where

---

Paper 61-5, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted March 18, 1960; made available for printing October 6, 1960.

G. C. TEMES is with the Northern Electric Company, Ltd., Ottawa, Ont., Canada.

The author wishes to express his appreciation to A. N. Oldfield, who made the present work possible; to J. D. MacDonald, who took an active part in the design process; and to T. M. Symbalisty, who built and tested the filters. He also wishes to acknowledge the co-operation of William Hyndman and John Dolan of the Computing Devices of Canada, Ltd., Computing Centre.

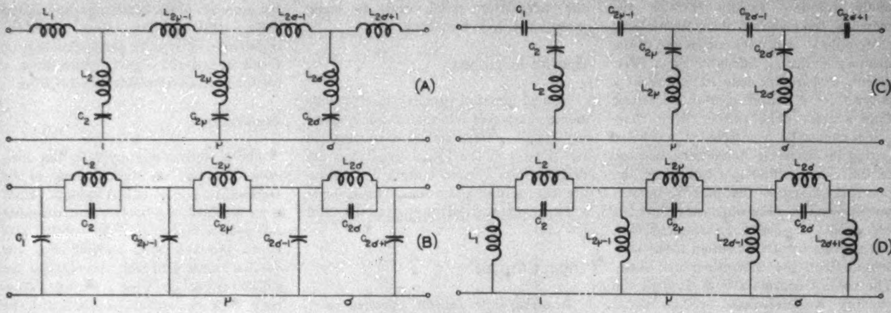


Fig. 1. Circuit diagram of general parameter filters

A—Low-pass mid-series
B—Low-pass mid-shunt
C—High-pass mid-series
D—High-pass mid-shunt

σ = number of attenuation poles at finite frequencies
$\mu = 1, 2, \ldots \sigma$
$\delta_0 = 1$
$\delta_1 = m_1 + m_2 + \ldots + m_\sigma$
$\delta_2 = m_1 m_2 + m_1 m_3 + \ldots + m_{\sigma-1} m_\sigma$
$\ldots$
$\delta_\sigma = m_1 m_2 \ldots m_\sigma$
$m_i = \Omega_{\infty i}^{-1}\sqrt{\Omega_{\infty i}^2 - 1}$
$\Omega_{\infty 1}, \Omega_{\infty 2} \ldots \Omega_{\infty i} \ldots \Omega_{\infty \sigma}$ = (normalized) pole frequencies

Also

$$s_i = \frac{1}{\Omega_{\infty i}^2}$$

Simplified formulas for $b_{2\mu}$ up to $\sigma = 4$ (four attenuation poles) can be found in the literature.[3]

After $\epsilon^{2\alpha}(p)$ has been obtained, the design follows the insertion-loss filter synthesis procedure as given by Darlington.[1] Substituting

$$\sqrt{\epsilon^{2\alpha_p} - 1}\ \sinh \theta_I = \Omega\,\frac{B(\Omega)}{P(\Omega)}$$

where B and P are even polynomials in Ω, the expression $P + pB$ can be factorized as

$$P + pB = (P_1 + pB_1)(P_2 - pB_2)$$

where P_1, P_2, B_1, B_2 are also even polynomials and $P_1 + pB_1$ contains all roots of $P + pB$ with negative real parts. The frequency and impedance normalized lattice elements of the desired filter can now be calculated from

$$Z_A = \frac{pB_1}{P_1}$$

$$Z_B = \frac{P_2}{pB_2}$$

If, as usually is the case, a ladder rather than a lattice configuration is desired, Table II of reference 1 can be used (up to $\sigma = 3$) to obtain the ladder coefficients.

Design Procedure for 1-, 2-, and 3-Pole General Parameter Filters

Calculations based on the theory as described above lead to the following design procedure for 1-, 2-, and 3-pole general parameter filters.

SINGLE-POLE (ONE-SECTION) GENERAL PARAMETER FILTERS

1. Given

f_p = pass-band limit frequency, cycles per second (cps)*
f_∞ = pole frequency, cps
$(R_{gen} = R_{load})$ = terminating impedances, ohms
α_p = pass-band ripple, db

2. The normalized pole frequency:

$$\Omega_\infty = \frac{f_\infty}{f_p} \quad \text{for low pass}$$

$$\Omega_\infty = \frac{f_p}{f_\infty} \quad \text{for high pass}$$

3. Steps 3–6 give design parameters previously mentioned.

$$m = \frac{\sqrt{\Omega_\infty^2 - 1}}{\Omega_\infty}$$

4.

$$s = \frac{1}{\Omega_\infty^2}$$

5.

$$b_0 = 1 + 2m$$

$$b_2 = (1 + m)^2$$

6.

$$H = \sqrt{10^{0.1\alpha_p} - 1}$$

7. To factorize $P + pB$, the $P(p) + pB(p) = 0$ equation has to be solved. This is equivalent to

$$p^3 + \frac{s}{b_2 H}p^2 + \frac{b_0}{b_2}p + \frac{1}{b_2 H} = 0$$

The solution of the equation gives a negative real root $(-r_1)$ and a pair of conjugate complex roots with positive real parts $(r_2 \pm j i_2)$. Here $r_1, r_2, i_2 > 0$.

8. The ladder coefficients:

$$a_1 = a_3 = \frac{1}{r_1}$$

$$a_2 = \frac{4r_2}{r_2^2 + i_2^2}$$

9. The impedance units of the normalized network:

$$C_0 = \frac{1}{2\pi f_p R_{gen}}$$

$$L_0 = R_{gen}^2 C_0$$

10. Finally, the denormalized element values can now be obtained:

Low-pass mid-series network [Fig. 1(A)]:

$$L_1 = L_3 = a_1 L_0$$

$$L_2 = \frac{s}{a_2}\ L_0$$

$$C_2 = a_2 C_0$$

Low-pass mid-shunt network [Fig. 1(B)]:

$$C_1 = C_3 = a_1 C_0$$

$$C_2 = \frac{s}{a_2}\ C_0$$

$$L_2 = a_2 L_0$$

High-pass mid-series [Fig. 1(C)]:

---

* If predistortion is to be avoided, a proper safety margin must be left in the f_p and α_p values used, to compensate for the effects of dissipation in the filter elements.

$$C_1=C_3=\frac{C_0}{a_1}$$

$$C_2=\frac{a_2}{z}C_0$$

$$L_2=\frac{L_0}{a_2}$$

High-pass mid-shunt [Fig. 1(D)]:

$$L_1=L_3=\frac{L_0}{a_1}$$

$$L_2=\frac{a_2}{z}L_0$$

$$C_2=\frac{C_0}{a_2}$$

TWO-POLE (TWO-SECTION) FILTERS

1. Input data:

f_p = pass-band limit
α_p = pass-band ripple
$f_{\infty 1}, f_{\infty 2}$ = pole frequencies
$(R_{gen}=R_{load})$ = terminations

2.

$$\Omega_{\infty i}=\frac{f_{\infty i}}{f_p} \qquad i=1,2 \qquad \text{for low pass}$$

$$\Omega_{\infty i}=\frac{f_p}{f_{\infty i}} \qquad i=1,2 \qquad \text{for high pass}$$

3.

$$m_i=\frac{\sqrt{\Omega_{\infty i}^2-1}}{\Omega_{\infty i}}$$

4.

$$z_i=\frac{1}{\Omega_{\infty i}^2}$$

5.

$$\delta_1=m_1+m_2$$

$$\delta_2=m_1m_2$$

6.

$$\sqrt{e^{2\alpha_p(N)}-1}=\sqrt{10^{0.1\alpha_p(\text{db})}-1}=H$$

7.

$$b_0=1+2\delta_1$$

$$b_4=(1+\delta_1+\delta_2)^2$$

$$b_2=b_0+b_4-\delta_2^2$$

8. The equation $P(p)+pB(p)=0$ is now equivalent to

$$p^5+\frac{z_1z_2}{b_4H}p^4+\frac{b_2}{b_4}p^3+\frac{z_1+z_2}{b_4H}p^2+\frac{b_0}{b_4}p+\frac{1}{b_4H}=0$$

The solution of this equation gives a negative real root $(-r_1)$ and two pairs of conjugate complex roots, one pair with negative real parts $(-r_2\pm ji_2)$ and one pair with positive real parts $(r_4\pm ji_4)$. Here

$$r_1, r_2, r_4, i_2, i_4>0$$

9. The parameters of the lattice network (reference 1, Table II) are

$$\gamma_1=\frac{1}{r_1(r_2^2+i_2^2)}$$

$$\gamma_2=\gamma_1(2r_1r_2+r_2^2+i_2^2)$$

$$\gamma_3=\frac{1}{2r_4}$$

$$\gamma_4=\gamma_3(r_4^2+i_4^2)$$

$$\beta_1=\gamma_1(r_1+2r_2)$$

10. The ladder coefficients can now be calculated:

$$a_1=\gamma_2-\gamma_4z_1$$

$$a_3=\gamma_4(2\beta_1-z_1-z_2)$$

$$a_2=2\,\frac{\beta_1-z_1}{a_3}$$

$$a_4=2\,\frac{\beta_1-z_2}{a_3}$$

$$a_5=\gamma_2-\gamma_4z_2$$

11. The impedance units of the normalized network are

$$C_0=\frac{1}{2\pi f_pR_{gen}}$$

$$L_0=R_{gen}^2C_0$$

12. The denormalized circuit element values can now be obtained.

Low-pass mid-series filter [Fig. 1(A)]:

$$L_1=a_1L_0$$

$$L_2=\frac{z_1}{a_2}L_0 \qquad C_2=a_2C_0$$

$$L_3=a_3L_0$$

$$L_4=\frac{z_2}{a_4}L_0 \qquad C_4=a_4C_0$$

$$L_5=a_5L_0$$

Low-pass mid-shunt filter [Fig. 1(B)]:

$$C_1=a_1C_0$$

$$C_2=\frac{z_1}{a_2}C_0 \qquad L_2=a_2L_0$$

$$C_3=a_3C_0$$

$$C_4=\frac{z_2}{a_4}C_0 \qquad L_4=a_4L_0$$

$$C_5=a_5C_0$$

High-pass mid-series filter[Fig. 1(C)]:

$$C_1=\frac{C_0}{a_1}$$

$$C_2=\frac{a_2}{z_1}C_0 \qquad L_2=\frac{L_0}{a_2}$$

$$C_3=\frac{C_0}{a_3}$$

$$C_4=\frac{a_4}{z_2}C_0 \qquad L_4=\frac{L_0}{a_4}$$

$$C_5=\frac{C_0}{a_5}$$

High-pass mid-shunt filter [Fig. 1(D)]:

$$L_1=\frac{L_0}{a_1}$$

$$L_2=\frac{a_2}{z_1}L_0 \qquad C_2=\frac{C_0}{a_2}$$

$$L_3=\frac{L_0}{a_3}$$

$$L_4=\frac{a_4}{z_2}L_0 \qquad C_4=\frac{C_0}{a_4}$$

$$L_5=\frac{L_0}{a_5}$$

THREE-POLE (THREE-SECTION) FILTERS

1. Input data:

f_p = pass-band limit frequency
α_p = pass-band ripple
$f_{\infty 1}, f_{\infty 2}, f_{\infty 3}$ = pole frequencies
$(R_{gen}=R_{load})$ = terminations

2.

$$\Omega_{\infty i}=\frac{f_{\infty i}}{f_p} \qquad i=1,2,3 \qquad \text{for low pass}$$

$$\Omega_{\infty i}=\frac{f_p}{f_{\infty i}} \qquad \text{for high pass}$$

3.

$$m_i=\frac{\sqrt{\Omega_{\infty i}^2-1}}{\Omega_{\infty i}}$$

4.

$$z_i=\frac{1}{\Omega_{\infty i}^2}$$

5.

$$\delta_1=m_1+m_2+m_3$$

$$\delta_2=m_1m_2+m_1m_3+m_2m_3$$

$$\delta_3=m_1m_2m_3$$

6.

$$b_0=1+2\delta_1$$

$$b_2=3b_0+2(\delta_1\delta_2+\delta_2+\delta_3)+\delta_1^2$$

$$b_6=(1+\delta_1+\delta_2+\delta_3)^2$$

$$b_4=b_2+b_6-b_0-\delta_2^2$$

7.

$$H=\sqrt{10^{0.1\alpha_p(\text{db})}-1}$$

8.

$$p^7+\frac{z_1z_2z_3}{b_6H}p^6+\frac{b_4}{b_6}p^5+\frac{z_1z_2+z_1z_3+z_2z_3}{b_6H}p^4+\frac{b_2}{b_6}p^3+\frac{z_1+z_2+z_3}{b_6H}p^2+\frac{b_0}{b_6}p+\frac{1}{b_6H}=0$$

The roots with negative real parts:

Table I. Design Time for 2-Pole General Parameter Filter

Problem	Time Using Desk Calculator, Man-Days	Time Using Digital Computer, Machine-Minutes*
First	3	3 man-days+ 4 machine-hours
Second	3	15
Third	3	15
Fourth	3	15
Fifth	3	15
Total	15	3 man-days+ 5 machine-hours

* Except where otherwise indicated.

$-r_1, -r_2 \pm j i_2$

The roots with positive real parts:

$r_3 \pm j i_3, r_4 \pm j i_4,$

where again all $r_k, i_k > 0$.

9.

$p_k^2 = r_k^2 + i_k^2 \qquad k = 1, 2, 3, 4$

10. The lattice parameters:

$\gamma_1 = (r_1 p_2^2)^{-1} \qquad \beta_1 = (r_1 + 2r_2)\gamma_1$

$\gamma_2 = (2r_1 r_2 + p_2^2)\gamma_1$

$\gamma_3 = [2(r_3 p_4^2 + r_4 p_3^2)]^{-1} \qquad \beta_2 = 2(r_3 + r_4)\gamma_3$

$\gamma_4 = p_3^2 p_4^2 \gamma_3$

11. Introducing some new design parameters:

$T_1 = (\beta_1 - z_1)(\beta_2 - z_2) + (\beta_2 - z_1)(\beta_1 - z_2)$

$T_2 = 2(\beta_1 - z_1)(\beta_1 - z_2)(\beta_2 - z_1) + (z_1 - z_2)T_1$

$T_3 = 2(\beta_1 - z_1)(\beta_1 - z_2)(\beta_2 - z_3) + (z_3 - z_1)T_1$

$$T_4 = \frac{\gamma_4}{\beta_1 - \beta_2}$$

12. The ladder coefficients:

$$a_1 = \frac{\gamma_1 - \gamma_2 z_1}{\beta_1 - z_1} \qquad a_2 = \frac{2(\beta_1 - z_1)(\beta_2 - z_1)}{T_2 T_4}$$

$$a_3 = \frac{T_2 T_4(\beta_1 - z_2)}{T_1} \qquad a_4 = \frac{2(\beta_1 - z_2)(\beta_2 - z_2)T_1^2}{(\beta_1 - \beta_2)T_2 T_3 T_4}$$

$$a_5 = \frac{T_3 T_4(\beta_1 - z_1)}{T_1} \qquad a_6 = \frac{2(\beta_1 - z_3)(\beta_2 - z_3)}{T_3 T_4}$$

$$a_7 = \frac{\gamma_1 - \gamma_2 z_3}{\beta_1 - z_3}$$

13. The impedance units:

$$C_0 = \frac{1}{2\pi f_p R_{gen}} \qquad L_0 = R_{gen}^2 C_0$$

14. The circuit element values can now be obtained.

Low-pass mid-series [Fig. 1(A)]:

$L_{2\mu \pm 1} = a_{2\mu \pm 1} L_0 \qquad \mu = 1, 2, 3$

$$L_{2\mu} = \frac{z_\mu}{a_{2\mu}} L_0$$

$C_{2\mu} = a_{2\mu} C_0$

Low-pass mid-shunt [Fig. 1(B)]:

$C_{2\mu \pm 1} = a_{2\mu \pm 1} C_0 \qquad \mu = 1, 2, 3$

$$C_{2\mu} = \frac{z_\mu}{a_{2\mu}} C_0$$

$L_{2\mu} = a_{2\mu} L_0$

High-pass mid-series [Fig. 1(C)]:

$$C_{2\mu \pm 1} = \frac{C_0}{a_{2\mu \pm 1}} \qquad \mu = 1, 2, 3$$

$$C_{2\mu} = \frac{a_{2\mu}}{z_\mu} C_0$$

$$L_{2\mu} = \frac{L_0}{a_{2\mu}}$$

Table II. Data Fed to and Obtained from the Computer in Course of the Synthesis and Checking of a Pilot Cut-off Filter

A—Synthesis of Filter

	s High Pass			
Input Data	0511600//	s		
	1600	112000/	s 56.112	f_p (observe margin)
	1602	103500/	s 56.1035	$f_{\infty 1}$
	1604	104500/	s 56.1045	$f_{\infty 2}$
	1606	600/	s 53.6	R
	1608	/1	s 50.1	α_p (observe margin)
	1610	300/	s 53.3	Q
	1612			
	0691500//	s		
Roots	−49.224916743506	51.101840356735		$-r_2 \pm j i_2$
	50.150688054223	51.103458996716		$r_4 \pm j i_4$
	−51.163547098302			$-r_1$
Network	C_1 43.107277267614	C_3 42.255160903292		C_5 43.147442623355
		C_2 41.814193025888	C_4 41.701483014023	
		L_2 48.290423788686	L_4 48.330666747852	

B—Frequency Response of Filter

Frequency	Attenuation, Db	Frequency	Attenuation, Db
101500	29.4959364	151500	0.0649944
102500	37.7369481	152500	0.0671135
103500	56.6351103	153500	0.0691837
104500	56.3721108	154500	0.0711998
105500	33.7390684	155500	0.0731577
106500	22.7851690	156500	0.0750542
107500	15.1671517	157500	0.0768870
108500	9.2268557	158500	0.0786543
109500	4.7798365	159500	0.0803550
110500	2.0670149	160500	0.0819886
111500	0.8487026	161500	0.0835548
112500	0.4292434	162500	0.0850539
113500	0.3093187	163500	0.0864862
114500	0.2784893	164500	0.0878526
115500	0.2666070	165500	0.0891539
116500	0.2539534	166500	0.0903912
117500	0.2369388	167500	0.0915659
118500	0.2165129	168500	0.0926792
119500	0.1944567	169500	0.0937325
120500	0.1723083	170500	0.0947274
121500	0.1511478	171500	0.0956655
122500	0.1316396	172500	0.0965483
123500	0.1141333	173500	0.0973775
124500	0.0987597	174500	0.0981546
125500	0.0855071	175500	0.0988814
126500	0.0742758	176500	0.0995594
127500	0.0649165	177500	0.1001902
128500	0.0572562	178500	0.1007755
129500	0.0511144	179500	0.1013169
130500	0.0463137	180500	0.1018158
131500	0.0426860	181500	0.1022738
132500	0.0400758	182500	0.1026923
133500	0.0383418	183500	0.1030730
134500	0.0373568	184500	0.1034170
135500	0.0370081	185500	0.1037260
136500	0.0371957	186500	0.1040011
137500	0.0378321	187500	0.1042437
138500	0.0388409	188500	0.1044551
139500	0.0401555	189500	0.1046365
140500	0.0417181	190500	0.1047891
141500	0.0434789	191500	0.1049140
142500	0.0453949	192500	0.1050125
143500	0.0474291	193500	0.1050855
144500	0.0495500	194500	0.1051341
145500	0.0517303	195500	0.1051594
146500	0.0539472	196500	0.1051622
147500	0.0561811	197500	0.1051436
148500	0.0584155	198500	0.1051045
149500	0.0606366	199500	0.1050458
150500	0.0628327	200500	0.1049682

High-pass mid-shunt [Fig. 1(D)]:

$$L_{2\mu \pm 1} = \frac{L_0}{a_{2\mu \pm 1}} \qquad \mu = 1, 2, 3$$

$$L_{2\mu} = \frac{a_{2\mu}}{z_\mu} L_0$$

$$C_{2\mu} = \frac{C_0}{a_{2\mu}}$$

By using the familiar frequency transformations, the results given above can of course be directly applied to the design of band-pass and band-stop filters with frequency-symmetrical attenuation characteristics.

Digital Computers as Design Aids

As the step-by-step procedure given above shows, the design of general parameter filters, while straightforward in principle, involves an unusual amount of numerical calculations accurate to about ten significant figures. The factorization of $P+pB$, for example, necessitates the solution of $P+pB=0$, which is a third-degree equation for $\sigma=1$, fifth-degree equation for $\sigma=2$, and seventh-degree equation for $\sigma=3$ with predominately complex roots. The computations preceding and following the solution of the equation involve some 80 elementary operations for $\sigma=2$, some 160 operations for $\sigma=3$. Obviously, the fastest and most economical way of carrying out the synthesis is by using a digital computer.

The use of computers in solving network problems is treated in numerous recent publications.[7–13] While preparing and carrying out a program for the design of one filter is usually more expensive than doing the same design on a desk calculator, the program is retained on punched tape or cards and could be re-used for any number of problems of the same type. The costs are equalized after two or three problems and thereafter the method using the computer is much cheaper. For the design problem described above and for $\sigma=2$ the necessary time expenditure is given in Table I.

As the results of Table I show, even for problems occurring once or twice a year, the use of a digital computer is economical. For certain problems which occur more frequently, such as the calculation of the frequency response of ladder networks, frequency transformations and denormalizations, finding roots of polynomials, etc., the use of a digital computer, resulting in an extremely radical drop in costs, is imperative.

The computer programs prepared for the synthesis of 1-, 2-, and 3-pole general parameter filters proved to be extremely useful in the practical carrier equipment design work of this laboratory.

The programs follow closely the steps given above; in fact, the form and order of the steps have been shaped to match the operation of the digital computer. They consist of four parts:

1. Acceptance of the input data; finding the coefficients in the polynomial equation.
2. Solving the equation and separating the roots with negative and positive real parts.
3. Synthesis of the network from the roots.
4. Assuming realistic Q's for the coils, the frequency response of the dissipative filter can be computed and checked against the original requirements.

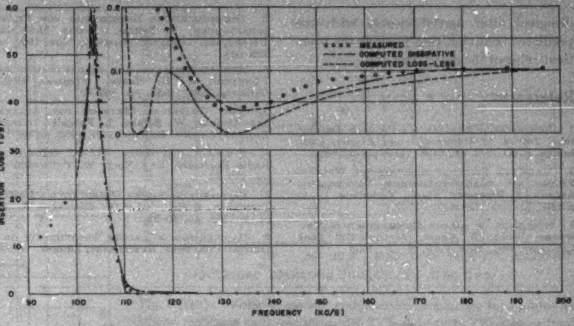


Fig. 2. Computed and measured frequency response of a general parameter filter

The roots of the equation (part 2) are obtained by Milne's quadratic factoring method.[13]

To perform part 4, a general-purpose checking routine is used. This computes the frequency response of ladder networks automatically. It is based on the theory of simple continuants.[14]

After the input data (f_p, α_p, $f_{\infty 1}$, Q, and $R_{\rm ter}$) have been fed in, the computer performs all design steps automatically. It types out the roots (which give a good check of the calculations) and the capacitance and inductance values of the synthesized network without human assistance. The frequency response of the filter could then immediately be obtained from the checking routine.

Design Example

A simple design example illustrates the practical use of the method. For a carrier equipment application a pilot cut-off filter was required to satisfy the following requirements:

Filter type: high pass
Pilot (pole) frequency: 104 kc
Minimum attenuation at pole frequency: 40 db
Pass-band limit: 116 kc
Maximum pass-band ripple: 0.3 db
Terminations: 600 ohms
Minimum Q of the available coils: 300

To minimize the number of coils in the network, the mid-series configuration [Fig. 1(C)] must be chosen.

Even though the design requirements as given above seem to indicate a single-pole filter, the finite Q of the available coils and the tolerance in element values made it necessary to establish two attenuation poles at $0.964\,f_\infty$ and $1.037\,f_\infty$.

The data presented to and received from a Bendix G-15 computer in course of the synthesis and the checking of the filter are shown in Table II. The roots and the network element values were typed out in floating-decimal-point notation.

As will be observed, there has been a safety margin left between the actual pass-band limit and the value fed into the computer (112 kc instead of 116 kc). There is also a safety margin in α_p (0.3 db→0.1 db). This is to compensate for the effect of parasitic dissipation. It is also of interest to notice that the type-out of element values follows closely the actual network configuration [compare Fig. 1(C)]. Fig. 2 gives the computed and measured frequency response of the filter.

Conclusions

Following a short summary of the insertion loss theory of symmetrical general parameter filters, a step-by-step design procedure has been presented for 1-, 2-, and 3-pole filters. The procedure comprised the solution of a higher degree polynomial equation and some elementary calculations. It has been shown that because of the volume of numerical computations involved, the use of a digital computer provides the only economical way of synthesis. For this reason, the design procedure was given in a form which lends itself readily to computer programming.

The design and checking of a pilot cut-off filter, carried out without human assistance, illustrated the effectiveness of this synthesis method. The measured frequency response of the computer-designed filter agreed closely with the predicted behavior, thus proving the practical value of the design method.

References

1. Synthesis of Reactance 4-Poles, **S. Darlington.** *Journal of Mathematics and Physics*, Cambridge, Mass., vol. 18, Sept. 1939, pp. 257–353.

2. Über den Entwurf elektrischer Wellen-Filter mit vorgeschriebenem Betriebsverhalten, **E. Rumpelt.** *Thesis*, Technische Hochschule, Munich, Germany, 1947.

3. Synthesis of Linear Communication Networks (book), **W. Cauer.** McGraw-Hill Book Company, Inc., New York, N. Y., 1958.

4. Die numerische Berechnung von Filterschaltungen, **V. Feltzer.** *Archiv Elektrischen Ubertragung*, Stuttgart, Germany, Nov. 1951, pp. 499–508.

5. Zur Aufstellung der charakteristischen Funktion, **K. H. Haase.** *Frequenz*, Berlin, Germany, May 1952, pp. 168–79.

6. New Graphical Methods for Analysis and Design, **W. Saraga, L. Fosgate.** *Wireless Engineer*, London, England, Mar. 1952, pp. 68–72.

7. *WESCON Convention Record*, Institute of Radio Engineers, New York, N. Y., session 26, pp. 133–53.

8. Predistorted Filter Design with a Digital Computer, **Geffe.** *WESCON Convention Record*, pt. 2, 1958, pp. 10–22.

9. Calculation of Ladder Coefficients on a Digital Computer, **Skwirzinsky, Zdunek.** *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, Dec. 1958, pp. 328–33.

10. On the Design of Filters by Synthesis, **Saal, Ulbrich.** *Ibid.*, pp. 284–327.

11. Katalogisierte Filter, **Glowatzky.** *NTZ-Nachrichtentechnische Zeitschrift*, Brunswick, Germany, vol. 9, Nov. 1956, pp. 508–13.

12. On the Tabulation of Insertion Loss Low-Pass Chain Matrix Coefficients and Network Element Values, **Bedrosian, Luke, Putschi.** *Proceedings*, National Electronics Conference, Chicago, Ill., vol. II, 1955, pp. 697–717.

13. Numerical Calculus (book), **W. E. Milne.** Princeton University Press, Princeton, N. J., 1949.

14. Principles of Circuit Synthesis (book), **Kuh, Pedersen.** McGraw-Hill Book Company, Inc., 1959, app. 1.

A Fatigue-Free Silicon Device Structure

W. B. GREEN
NONMEMBER AIEE

GERMANIUM DEVICES, when correctly used, are capable of essentially infinite life. It was commonly assumed in early higher temperature semiconductor work that silicon devices and those of materials, such as silicon carbide and III-V compounds would have similar life expectancy. That this is not realized in soft-soldered silicon devices is now well established. Evidence is presented to show that this limitation in life is due to fatigue rupture of the soft-soldered joints when they are subjected to cyclic thermal stress. A new hard-soldered structure is described which has been proved by long experience to eliminate fatigue failure completely.

The Problem

Historically, the first large-area silicon devices were alloy junction rectifying cells. These rectifying cells offered substantial advantages over earlier devices such as those of selenium and germanium. These advantages took one or more of the forms of low cost, high efficiency, low forward voltage, light weight, small size, high inverse voltage, or the ability to operate in higher ambient temperatures. Among the early cell applications, a number of installations were encountered where unusually high failure rates were recorded. A common feature of these installations was that there were present a large number of "off-on" operations each day.

To understand the problem thus revealed, a test cycle was designed to provide the most difficult conditions normally encountered in the field. A high-current d-c supply and a high-voltage high-resistance d-c supply were both connected across a bridge of silicon rectifying cells mounted on copper plates. The current supply was alternately opened and closed by means of a cam-controlled timer. The time cycle chosen was $3^1/_2$ minutes on, and $6^1/_2$ minutes off. The value of current used was that necessary to heat the junction to a specified temperature in the time allowed. The voltage applied was such that its peak value was not greater than the rated piv (peak inverse voltage) of the cells. Fig. 1 shows a plot of a typical set of test data. This test cycle has the advantage that if either the reverse voltage or thermal resistance of the cells deteriorates substantially, failure will occur. Failures did occur, and in an alarmingly small number of cycles. These failures lead first to a careful re-examination of the effect of current and voltage pulses on cell operation.

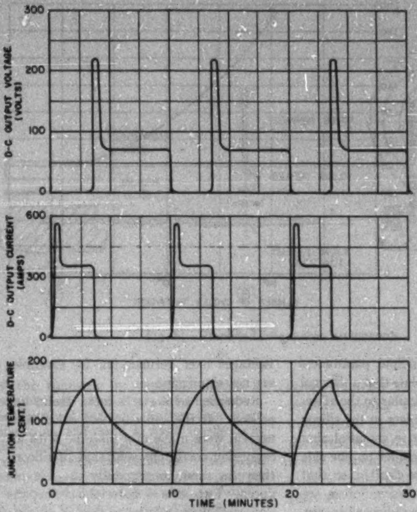
The maximum overload current allowable as a function of the time for which the overload operates is seen in Fig. 2. This allowable overload is for a cell carrying rated current and operating at rated temperature at the time of application of the overload. This curve is a statement of the fact that operation of these cells can be impaired in the forward direction by thermal energy alone. As is expected, shorter operating times lead to higher permissible overloads. The amount of thermal energy required for permanent damage to the cell is not quite constant over the whole range of times, since for short overloads the heat must largely be dissipated in the vicinity of the junction, while for longer times the case and cooling fin may assist the dissipation. In the test cycle employed here, as in many of the problem installations, all of the operating conditions are so well defined that failure caused by current overload could be ruled out.

Overload in the reverse direction, wherein the overload takes the form of a voltage higher than the rated piv of the junction, presents a much different problem than forward current overload. A junction exposed to this form of overload will break down; in this condition, the only real limitation to the amount of current that flows is the impedance of the voltage source and the load. Thermal destruction of the junction in the reverse direction will occur in comparable times and with comparable amounts of energy per unit area to that required in the forward direction. In order that the junction remain in a condition of breakdown, a high voltage must be present. Therefore, much smaller currents are required in the inverse direction than in the forward direction for thermal destruction. The fact that breakdown occurs essentially at a point, further indicates that destruction currents are small. Under these conditions, destruction will ordinarily occur in a very short time. It is possible in principle to plot a series of curves showing permissible overvoltage as a function of time for various source impedances. Such data are not now available; if they were, they could probably best be summarized by the statement that for most industrial applications a rectifying cell will surely be destroyed if in any form a peak inverse voltage greater

---

Paper 61-24, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 25, 1960; made available for printing December 19, 1960.

W. B. Green is with the Westinghouse Electric Corporation, Youngwood, Pa.



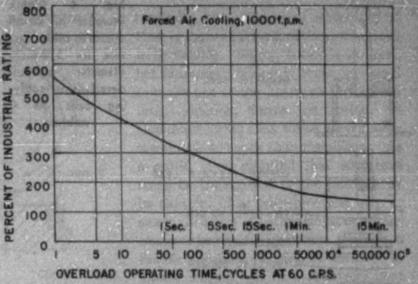


Fig. 1 (left). Voltage, current, and junction temperature for a bridge of silicon rectifying cells operating with 35% duty cycle 375-ampere load 1/2-inch-diameter junction

Fig. 2 (above). Maximum overload rating for 1/2-inch-diamete silicon junction starting from recommended industrial rating on 5x5x1/8-inch copper plate at 25 C ambient

than breakdown voltage is applied. It is ordinarily one of the most difficult problems of application of semiconductor rectifying cells to be sure that the magnitude and duration of voltage transients is known under all line and load conditions. Again, as in the case of forward current overload, measurement both of field and laboratory conditions ruled out failure due to reverse voltage breakdown.

Careful attention was given to failures where both voltage and current surges were limited and where failed units were promptly removed from load to prevent overheating. These defective units all had a common characteristic, partial or complete fracture of one or both of the soft-solder joints. When the fracture had progressed substantially through the solder joint, thermal resistance of the connection rose so as to allow the rectifier to fail due to excessive deterioration of reverse characteristics.

The construction of the rectifying cell used in the early parts of this work is seen in Fig. 3(A). The basic rectifying element consisted of a sandwich of molybdenum contacts, ohmic alloy, counter-electrode alloy, and silicon.[1] This element was soldered with a predominately lead soft solder to a copper case and flexible connection. Other commercial designs differed from this in the absence of one or both of the moly contacts, the composition of the alloys, or the composition of the soft solder; however, Fig. 3(B), which is a simplified model of this assembly useful for thermal considerations, was applicable to most conventional silicon rectifiers of that date. With only minor alteration, this model applied also to silicon transistors.

Using the cycle shown in Fig. 1 and by varying the short-circuit current, Fig. 4 was obtained showing the number of "on-off" cycles to solder fracture for 5/8-inch junction soft-soldered cells. Simply stated, these data state that the greater the temperature above ambient to which the junction rises, the fewer times it may be turned on and off before failure occurs due to fracture of the soft solder. Data collected on several size junctions over a range of junction temperatures may be summarized in the relation

$$ND^3\Delta T^3 = C^3 \qquad (1)$$

where N is the number of cycles to failure, D the junction diameter in inches, ΔT the change in junction temperature, and C a constant of value about 3,500. The lower temperature portion of the curve shown by dashed lines is added because germanium units cycled to this temperature apparently show little or no such deterioration. It is unreasonable to derate silicon junction temperature to 80 C (degrees centigrade), for then its principal advantage would be lost. Thus, while one may assume that a germanium device when conservatively rated and correctly applied will have practically infinite life, a silicon device with soft-soldered construction, properly rated and applied, had a finite life readily determined in advance. For many installations, this relatively short cyclic life was a serious limitation.

The Nature of Fatigue Failure

Metallurgical literature contains a number of references to mechanical failures of joints under cyclic conditions relevant in some aspects to these rectifying cell failures. For example, in the automotive industry, radiators with soft-soldered joints fail at the joints when subjected to cyclic stress.[2] For test purposes, in the latter, a radiator-like container is fabricated by soft soldering two hemispherical shells together. The container thus formed is alternately filled with steam under pressure and allowed to cool slightly with no steam pressure. The essential feature of the test is the cyclic stress placed across the solder at a high temperature. Number of cycles to failure of the joints was found to depend partially on soft-solder composition.

An even more pertinent series of tests were conducted on thermal stress cycling of austenitic stesel and resultant fatigue failure.[3] In this case, a steel bar was fastened at both ends and then heated over a range of temperature ΔT so as to generate internal stress in the bar due to constrained thermal expansion. It was shown that if N is the number of cycles to failure of the bar, then

$$N^{1/2}\Delta\epsilon_p = C \qquad (2)$$

where $\Delta\epsilon_p$ is the change in plastic strain in

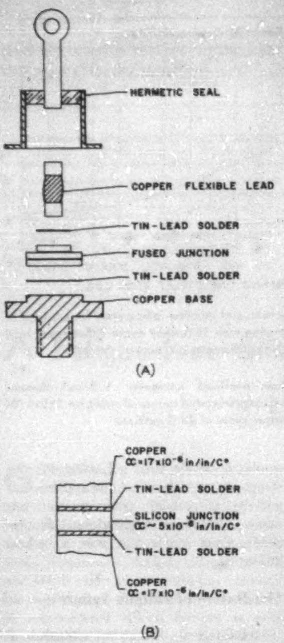


Fig. 3. A—Soft-solder silicon rectifying cell structure. B—Simplified thermal equivalent of this structure

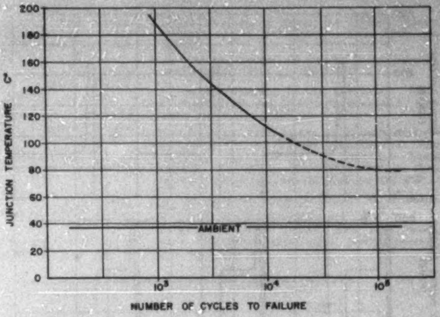


Fig. 4 (right). Number of on-off cycles to solder fracture for 5/8-inch-diameter soft-soldered silicon rectifying cells 38 C ambient

one cycle and C is a constant which depends on material, heat treatment, and cold work, among others. When the experimental bar contained either a joint between two different metals or a stress-raising defect, the failure occurred at the joint or defect. For the case of the joining of dissimilar metals, the joint was under a higher stress due to the difference of thermal expansion coefficients of the two metals. For the joint of dissimilar metals,

$$\Delta\epsilon_p = f_1(D,\Delta\alpha) + f_2(\alpha_1+\alpha_2)\Delta T \tag{3}$$

where D is the diameter of the bar and $\Delta\alpha$ is the difference of thermal expansion coefficients α_1 and α_2 of the two metals. The function f_1 relates to the stress at the joint due to the mismatched expansion coefficients while the function f_2 is the stress increment generated by the longitudinal thermal expansion and its consequent internal forces.

The experimental evidence, both with respect to the nature of the soft-solder failure and to the principal parameters controlling failure, suggests that a thermal stress induced fatigue similar to the aforementioned was the problem in the silicon cell. In such a cell, it is essential that the current-carrying member on one side of the semiconductor be flexible, so that direct tensile or compressive forces are excluded. However, as indicated in Fig. 3(B), there is a substantial difference in thermal expansion coefficient between the silicon element and the case and flexible connection. This $\Delta\alpha$ results in a $\Delta\epsilon$ which can easily be shown to be plastic by a simple calculation for the case of soft solders in the usual melting range. It follows that the stress at the periphery of the solder joint will increase with diameter. Then

$$\Delta\epsilon_p = f(\Delta\alpha, D)\Delta T \tag{4}$$

where $\Delta\alpha$ is essentially constant for all sorts of devices using the same semiconducting material. A reasonable relation for predicting fatigue failure in soft-soldered units is, then,

$$N^{1/2}\Delta T f(D) = C \tag{5}$$

The experimental data indicate a good fit for $f(D) = D$, the junction diameter, as stated in equation 1.

Fatigue-Resistant Components

Several methods of increasing the value of N are possible based on the previous theory. The values of either or all of $\Delta\epsilon_p$, $\Delta\alpha$, and ΔT may be reduced. Of these, a reduction in ΔT is least useful. For use as in power rectifiers, the primary advantage of a silicon rectifying cell over those of other semiconducting materials lies in the fact that allowable junction operating temperatures are high. If the junction temperature is sufficiently reduced to eliminate fatigue, the advantages over germanium, for example, are not significant.

Reduction of $\Delta\alpha$ is a more useful approach. If the cell case and other connection were made of a material with an expansion coefficient near that of silicon, then $\Delta\alpha$, and consequently $\Delta\epsilon_p$, is reduced. Two sorts of material have expansion coefficients similar to that of silicon: the high-nickel steels or kovar, and the refractory metals such as tungsten, molybdenum, and tantalum. While neither group of materials can be absolutely ruled out, the lower thermal conductivity of the former and difficulty of fabrication of the latter make their general use undesirable. To obtain optimum thermal and electrical characteristics in the finished product, a high-conductivity case material such as aluminum or copper is necessary.

The silicon disk with the fused alloys required to make the junctions may not be subjected to any tensile or bending stresses and can tolerate only modest compressive stress. Soft solders which do not transmit these stresses are eliminated by fatigue failure. Thus, a contact on each side of the device must be provided which has a coefficient of thermal expansion equal to or slightly greater than silicon. Tungsten, molybdenum, or tantalum, or suitable alloys of these and other materials are usually chosen. Since a copper case is usually desired, $\Delta\alpha$ is unchanged, but $\Delta\epsilon_p$ may be eliminated in several ways.

Three somewhat distinct procedures are now available for achieving this purpose; one example is cited from each.

1. A disk of contact material may be used which is so thick and so strong that if directly hard soldered to copper, most of the resulting strain will occur in the copper. An example is shown in Fig. 5(A). Here the tungsten contact would have a diameter-to-thickness ratio of the order of 5 to 1. This is adequate to assure that no bending

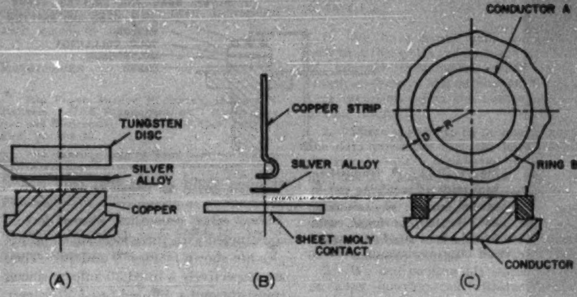


Fig. 5. Simplified fatigue-free components

of the tungsten by bimetallic action is possible. If the silver alloy makes a sufficiently strong joint, the horizontal stress due to $\Delta\alpha$ will result in small tensile $\Delta\epsilon$ in the tungsten and large compressive $\Delta\epsilon$ in the copper. Strain calculations for this structure show that $\Delta\epsilon$ in the copper is entirely elastic, and hence not subject to fatigue. A logical extension of this structure would be to place materials of intermediate expansion coefficients between the copper and tungsten, or even to fabricate a sintered structure incorporating a smooth composition transition from pure copper to pure tungsten. These latter seem unnecessarily complicated in view of the success of the former.

2. For low-power rectifying cells or for most small transistors, some or all of the connections may be preferably made in a less cumbersome manner. Since the current is low, a thin copper strip is adequate for connections. A satisfactory fatigue-free connection may be achieved by welding or brazing this strip to a contact of molybdenum, tungsten, or tantalum, or in some cases iron or nickel which is fused on the device in its initial fabrication or in a subsequent lower temperature operation. If the diameter-to-thickness ratio of the contact is of the order of 20 to 1 and the contact-to-strip thickness ratio of the order of 5 to 1, a satisfactory result is achieved. This form of connection allows the contact to be made from punchings of sheet material. This connection is shown in Fig. 5(B).

3. For connections of cylindrical symmetry, it is possible to restrict the expansion of a conductor and hence use it directly as a contact. In Fig. 5(C), a conducting material A (such as copper) is surrounded by a ring of low-expansion material B (such as tungsten). If Young's modulus of B is substantially greater than that of A, and if $D \geq R$ then the thermal expansion coefficient of the combination will be nearly that of A.

The principle acting in all three of these cited examples is simply that if the joint between a refractory metal and a current-carrying connection is strong enough and of the right geometry, the strain from $\Delta\alpha$ will occur largely in the current-carrying member. Copper has a sufficiently high yield point that, whereas the stress was sufficient to lead to considerable plastic deformation in lead-tin solders, now only elastic deformation results.

The choice of alloys to make the indicated joints is governed by several considerations. Most important of these is that the alloy wet well the refractory metal chosen. The alloy must also be strong enough so that the stress in the joint does not result in plastic deformation of the alloy. The melting range of the alloy will be determined by the melting ranges of the alloys in the device; higher if a preassembly is contemplated, or equal if the contacts will be joined to the case or connections during the orignal device fusion. For the cell described in the following section, since the ohmic alloy is silver-lead-antimony, silver alloys were chosen which had a comparable melting range. Copper, gold, or aluminum should be equally satisfactory as an alloying material.

A Practical Rectifying Cell Structure

There are several considerations which dictate the design of a practical silicon rectifying cell. The most important of these is the need for one of the connections to be flexible. Silicon is brittle in itself, but aluminum, which is commonly used to make one of the junctions in a silicon fused junction, forms, on fusion, an aluminum silicon alloy which is even more susceptible to shock than the silicon. As a result, it is essential that minimum stress be transmitted to the junction by its capsule. The only practical way to accomplish this is by making one of the connections to the junction through a flexible member. A flexible member in most instances will be a relatively poor thermal connection either by virtue of its length or reduced cross section. It is necessary, therefore, that the heat generated in the junction be removed through one connection only, and that one be designed in some fashion for attachment to a cooling fin or other heat sink.

The construction of a rectifying cell in which these principles are applied is shown in Fig. 6. Two sizes of rectifying cells are made in the configuration shown, 1/2- and 5/8-inch-diameter junctions with ratings of about 150 and 225 amperes respectively. The components shown are rigorously cleaned and assembled with the aid of jigs and weights. The assembly is fused in an atmosphere furnace. A flexible copper lead is mechanically swaged into the copper tube at the top of the assembly. Further processing adds the components necessary to seal the unit hermetically. Fig. 7 shows units in various stages of assembly.

The test cycle used for evaluation of this cell is the same as that shown in Fig. 1. The only exceptions from the values given in Fig. 1 are that the diode used has a 1/2-inch junction and that the junction temperature rises to a peak of about 185 degrees. Cells made with conventional soft-solder techniques will have first failures after about 600 cycles with almost all units failing before 2,000 cycles. A hard-soldered assembly similar to that of Fig. 7 has already been tested through 90,000 cycles with no indication of

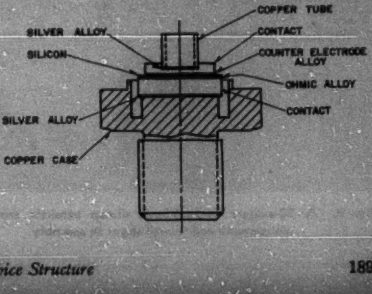


Fig. 6. A fatigue-free fused silicon rectifying cell structure

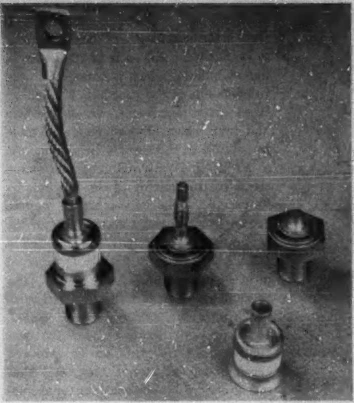


Fig. 7 (left). A fatigue-free hard-soldered rectifying cell in several stages of assembly: right to left, as fused, with flexible lead attached, encapsulation complete

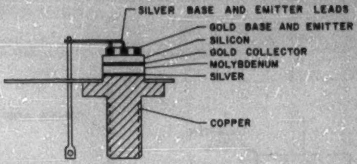


Fig. 8 (right, above). Cross section of a hard-soldered silicon power transistor

fatigue deterioration. In addition, nearly 3 years of field experience with cells of this type have yet to yield the first failed hard-soldered joint.

Fatigue-Free Transistor and Trinistor Structures

Recent advances in the field of alloy fusion of large-area devices have made possible the precise control required for the successful fusion of high-power silicon transistors. It is apparent from the experimental work presented that the geometric differences between transistors and rectifying cells are not significant with respect to fatigue failure. It is mandatory that all large-area devices of whatever nature be made with a hard-soldered fatigue-free structure.

The first transistor to contain this construction is shown, in section, in Fig. 8. It is a 5-ampere 250-volt device. Note that as in the previous rectifying cell, the primary thermal connection, to the collector, is from the silicon through a gold alloy, a thick molybdenum contact, and a silver alloy directly to the copper. In addition to its fatigue-free properties, this method of joining allows a reproducibility of thermal resistance not otherwise possible. The emitter and base contacts here are made of pure silver joined directly to the gold. In this case, the silver is a relatively weak, deformable member whose expansion characteristics at the point of contact are forced to conform to those of the silicon and molybdenum by a strong and relatively undeformable alloy. The high elastic limit of silver again assures that only elastic deformation is present.

Two other successful silicon devices constructed with these mechanical principles are shown in Figs. 9 and 10. They are respectively a fused 30-ampere silicon transistor and a diffused-fused 50-ampere silicon Trinistor switch.

Conclusions

1. A soft-soldered silicon device temperature which is cycled above about an 80 C junction temperature will fail in a number of cycles determined primarily by the temperature excursion in cycling and the diameter of the device.

2. The failure on cycling is due to fatigue rupture of the soft-solder joints and is caused by the difference of thermal expansion coefficient of the device and the heat sink. As a result, the time to failure is only slightly affected by the way the soft-solder joint is made.

3. If correct geometry is chosen, it is possible to hard solder refractory metal contacts directly to copper and achieve a joint which is free of plastic deformation.

4. Silicon rectifying cells incorporating such hard-soldered joints will not fail under cyclic conditions.

5. Silicon power transistors and Trinis-

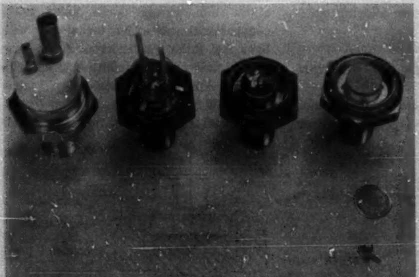


Fig. 9. A 30-ampere hard-soldered silicon transistor showing the components and several stages in assembly

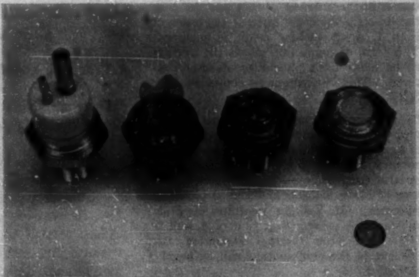


Fig. 10. A 50-ampere hard-soldered Trinistor switch showing component parts and several stages of assembly

tor switches incorporating hard-soldered connections have also been fabricated with excellent results.

References

1. The Fused Silicon Rectifier, Herbert W. Henkels. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, 1956 (Jan. 1957 section), pp. 733–46.

2. The Performance of Some Soft Solders at Elevated Temperatures and Pressures, H. E. Pattee, R. M. Evans. *Special Technical Publication 189*, American Society for Testing Materials, Philadelphia, Pa., 1956, p. 103.

3. The Problem of Thermal Stress Fatigue in Austenitic Steels, L. F. Coffin, Jr. *Special Technical Publication 165*, American Society for Testing Materials, 1954, p. 31.

◆

Discussion

D. K. Bisson (General Electric Company, Auburn, N. Y.): We can verify the conclusions on the thermal fatigue phenomenon in soft-soldered rectifiers. Our first experience was in 1956 when our Locomotive and Car Equipment Department observed early failures on a cycling test of our 50-ampere silicon cell. We also can substantiate the author's success with the hard-solder method with cells that have been on laboratory test since 1958, and with field experience since 1959. I have four comments or questions I would like to raise.

1. We have found with soft-soldered cells that the cycles to failure are a function of internal temperature gradients and length of cycle as well as junction temperature excursion. Has the author observed similar results?

2. There appears to be no question from the data presented that the hard-soldered design represents a substantial improvement in thermal fatigue resistance. Does the author expect that hard-soldered cells will never fatigue, or does he believe that a curve similar to Fig. 4 (with substantial increase in both cycles to failure and temperature excursion) could be expected?

3. We can confirm his statement that an alloy junction can withstand only a modest bending stress, usually accompanied by changes in reverse current. The thermal fatigue solution seems to be less difficult with diffused junctions, however. We have been able to deform a 1/2-inch-diameter junction assembly with an axial load to an 8-inch radius of curvature without fracturing. No detectable change of reverse current occurred up to actual fracture.

4. The author stated that when a refractory metal is bonded to copper, only elastic deformation results in the copper. We believe that plastic deformation does occur in the copper. Assuming a perfect bond, a tungsten-copper system will take the copper past its yield point with only a 150 C temperature change. While this can be expected to produce fatigue in the copper, we suspect that it is delayed for a very large number of cycles because the copper is not a thin layer and an almost infinite source of material exists near the interface in which slip can occur.

W. B. Green: In answering the discussion, I want first to emphasize that there have been no fatigue failures in the field or in the laboratory on the devices described in this paper. This undoubtedly covers many applications with total numbers of cycles much in excess of the published results. In answer to the specific questions:

1. Preliminary results did not show such a relationship, but we did not explore this in any detail.

2. and 4. Since there are no failures from which to gain information, one can only turn to analytical results. We examined the problems involved in accurately calculating the stress in the copper-tungsten joints and conclude that it is an almost impossible task. This is true due to the fact that for the actual geometry used bending due to bimetal effect must be incorporated. The one problem that could be reasonably solved was that of the joint between infinite lengths of cylindrical copper and tungsten. This solution indicated no failure within applicable conditions. Lacking a more concrete result than this, I suggest that field results justify considerable satisfaction.

Late Discussion

Discussion of "A Comparison of Digital Differential Analyzer and General Purpose Equipment in Guidance Systems," by M. M. Dickinson, published in Communication and Electronics, January 1961, pages 706–08.

Gerhard L. Hollander (Hughes Aircraft Company, Fullerton, Calif.): One key factor in the performance per unit cost of any computer is the speed of the circuitry. In this comparison, the DDA used 1-megacycle circuitry while the GP used 200-kc circuitry.

As circuit speed directly affects the cost, could the author compare the effect of using comparable circuitry in the DDA and GP? While the particular implementation may not have made the same clock rate desirable, a comparison with equivalent circuitry should yield useful conclusions for future designs.

To allow interested workers to delve more deeply into the assumptions underlying this paper, would the author reference more detailed descriptions of the two computers?

M. M. Dickinson (International Business Machines Corporation, Owega, N. Y.): Indeed, the information-processing rate is five times greater and the circuits are more complex in the DDA than in the GP. The circuit variation results from matching their performance to that of the storage elements in the two cases. The 1-megacycle circuits are matched to the magnetostrictive delay line but are too fast to work into the magnetic drum used in the GP computer. Since a magnetic drum is not available to match 1-megacycle circuitry and since the general purpose computer described is organized around a magnetic drum memory, it is not meaningful to consider a GP with the higher speed circuits. In the case of a 200-kc DDA, it would be necessary to operate the system on a more nearly parallel basis in order to realize the same computing capacity as given by the reported equipment. Such a reorganization would imply increase in cost factors by perhaps 200% to 400%.

A laboratory model of the reported digital differential analyzer has been built and is operative at the IBM FSD Space Guidance Center. A description of this computer is reported in "Final Report on an Air-to-Surface Missile-Guidance High-Speed Digital Differential Analyzer," *WADC-TR-59-651, IBM no. 60-911-6*, January 1960. A general purpose computer derived from the reported design is now in production at the same International Business Machines Corporation facility. Descriptive material for interested workers is available through E. J. Doyle, IBM FSD Space Guidance Center, Owego, N. Y.

Type T-12S-Tr Transistorized Short-Haul Carrier System

O. ISHIHARA
NONMEMBER AIEE

Y. YAMAMOTO
NONMEMBER AIEE

S. FUJITA
NONMEMBER AIEE

R. MURAKAMI
NONMEMBER AIEE

THE TOLL NETWORK in Japan has been expanded and improved through long-term projects, utilizing various types of new equipments. The main long-distance network has been built up of coaxial cable and microwave systems, which are used to connect offices larger than primary outlets. To connect offices smaller than primary outlets, a more economical system had to be considered, and with recent acceleration in conversion to the toll dialing system, it became imperative to increase rapidly the number of toll circuits between large cities and their surrounding satellites, as well as between the principal large cities.

Loaded cable has been used in the past to interconnect central and satellite cities. However, to provide economically the necessary number of circuits, it appeared desirable to convert the loaded-cable system to a multichannel carrier system by removing the loading coils or installing new toll cables.

To meet the demand for new circuits, the Nippon Telegraph and Telephone Public Corporation (NTT) decided in 1957 to develop a 12-channel 2-wire-basis short-haul carrier system. Development of this new system, designated *T-12S-Tr*, was undertaken in co-operation with the Nippon Electric Company (NEC) and a prototype model was completed in June 1958. The system was designed to satisfy the following requirements:

1. The cost of the system, when used to connect offices smaller than primary outlets, should be lower than that of the long-haul carrier system (coaxial cable system, 24-channel nonloaded cable system, etc.) and the loaded cable system. To reduce the cost, new circuit techniques and new components should be introduced, as for example, the single-stage modulation scheme using channel filters made up of ferrite-core coils and polystyrene capacitors.
2. The equipment should be fully transistorized, since this, together with the planned miniaturization of other components, would reduce power and floor-space requirements and allow for easier maintenance.
3. The system should be usable with existing (installed) toll cables.
4. The system should utilize the existing power supplies in the offices in which it would be installed.
5. Allowable noise for the system should conform to the noise-distributing standard for toll networks.

These requirements reveal that the *T-12S-Tr* was designed for both mechanical and electrical economy. Since the development of high-frequency transistors had been proceeding at a rapid pace in Japan, there seemed to be good prospects for their application to the carrier field. The adoption of transistors as well as the application of new circuit techniques for this system made it possible to attain the stated design objectives.

The first *T-12S-Tr* system was delivered and placed in operation in September 1958.[1] Since that time, the system has been operating very satisfactorily and the initially established objectives were achieved.

Summary of System

System Application

In general, the long-haul carrier system is designed for application to sections between offices larger than primary outlets, as shown in Fig. 1, and the short-haul carrier system is designed for use between offices smaller than primary outlets, such as between primary outlet and toll center, and between toll center and toll center. The application of the short-haul carrier system to sections between offices smaller than toll centers should also be considered. However, since the average length of these sections is about 10–15 km (kilometers), the use of the loaded cable system is, at present, more economical. Consequently this newly developed *T-12S-Tr* system is best used in sections between primary outlet and toll center, and between toll center and toll center. The distribution of the lengths of these sections is shown in Fig. 2. According to Fig. 2, it is reasonable that the short-haul carrier system should cover a maximum circuit length of 100 km.

System Performance

The performance of the *T-12S-Tr* system is as follows:

1. Transmission system:
   Separate-group 2-wire single-side-band transmission system.
2. Transmission channel capacity:
   12 telephone channels per system.
3. Cable and repeater spacing:

Cable Type	Capacitance, Microfarad per Km	Loss at 120 Kc, Decibels per Km	Repeater Spacing, Km
0.9 mm (millimeter) Awg (American wire gage) no. *19* star quad PEF toll cable	0.0385	3.5	7~12
0.9 mm (Awg no. *19*) star quad paper toll cable	0.0385	3.64	7~12
1.3 mm (Awg no. *16*) DM (Dieselhorst Martin) quad paper toll cable	0.0385	2.34	10.5~18

4. Line-frequency band:
   Low-group................12~60 kc
   High-group...............72~120 kc
5. Power supply voltage:
   Direct current −21 volt ±10% (in case of existing offices having −24-volt batteries, the power source is adjusted by the use of a voltage regulator).

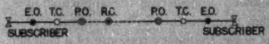


Fig. 1. Diagram of toll network

RC—Regional center
PO—Primary outlet
TC—Toll center
EO—End office

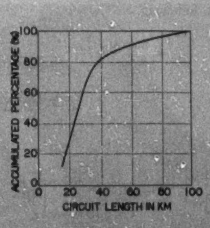


Fig. 2. Accumulated distribution curve of circuit length of short-haul carrier section

---

Paper 61-112, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 28, 1960; made available for printing December 20, 1960.

O. Ishihara and S. Fujita are with the Nippon Telegraph and Telephone Public Corporation, Tokyo, Japan. Y. Yamamoto and R. Murakami are with the Nippon Electric Company, Ltd., Kawasaki City, Japan.

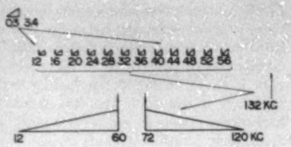


Fig. 3 (left). Frequency allocation

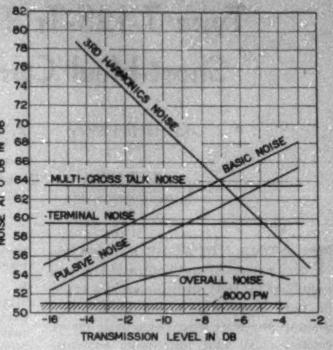


Fig. 4 (right). Transmission-level characteristics of various kinds of noise

6. Signaling system:
   3,850-cps (cycles per second) out-of-band idle-time sending system. The signal level should be 10 db (decibels) below the speech level.
7. Frequency allocation:
   As shown in Fig. 3.
8. Carrier frequency:
   The carrier frequency can be supplied to eight systems. The frequency accuracy is within $\pm 10^{-6}$. Automatic switchover from the operating system to an emergency system is provided.
9. Frequency response of a channel:
   Within 3/5 of the standard (these limits are illustrated with broken lines in Fig. 11) recommended by CCITT (International Telegraph and Telephone Consultive Committee) for direct international carrier circuits.
10. Over-all noise:
    8,000 pw (picowatts), $\mu\mu$w (micromicrowatts) (the psophometrically weighted noise power and 8,000 pw is equivalent to 34 dba (decibels adjusted) as measured with the Western Electric *2B* Noise Measuring Set) at 0-db reference point.

In designing this short-haul carrier, full consideration has been given to all details of the system and equipment to assure the required economic advantage, with the result that system and equipment are balanced not only from the economic but also from a performance-characteristics viewpoint. That is, in regard to system design it may reasonably be allowed, considering the applicable field, that requirements on such characteristics as noise, frequency response of a channel, and functions of carrier supply are relaxed. The allowance obtained can be utilized to simplify the circuitry and realize economical equipment. Thus system as well as each circuit are given various features, essentials of which are outlined in the following.

1. Transmission system. This system provides channels spaced at 4 kc, using the single-side-band method, and consequently does not require a wide transmission band.

2. Circuit over-all noise. Essential for the expected economic advantage is the amount of circuit noise that may be allowed in the system design. The most reasonable compromise between circuit-noise allowance and economic advantage is to allow over-all noise up to 8,000 pw (at 0-db point) for one 100-km link.

3. New 12×1 modulation system. The requirement on the circuit-noise allowance being relaxed, as described in the previous paragraph, it is possible to allow relatively high noise for the terminal equipment; i.e., the 12×1 direct modulation system used with channel filters of ferrite-core coils and polystyrene condensers can be adopted. With the use of this direct modulation system, the pregroup stage required in the pregroup modulation system (3×4 system) may be eliminated. This is a great economical advantage.

4. Signaling system. The signaling system used in this equipment is a 3.85-kc out-of-band single-frequency system which has the following characteristics:

(*a*) Out-of-band voice-frequency signaling system. Employing the 12×1 modulation system described previously, each channel-modulator input and voice-frequency amplifier output is connected to a ringer which transmits or receives the signaling frequency. It is then possible to simplify the signaling-frequency supply circuits, which need to supply only one voice frequency.

(*b*) High transmission level of signaling frequency. The circuit-noise allowance of up to 8,000 pw permits the system to operate with a signal-transmission level only 10 db below the reference level, which allows for a simplified ringer circuit. (With the long-haul system in Japan, the signal level is set 15 db below the reference level.)

(*c*) Adoption of reed relay. Since the circuit bundle in the short-haul carrier system is small, unlike that in the long-haul system, no back contact of ringer-output relay is required. Thus, instead of the polarized relay in the long-haul system the reed relay may be adopted, for economy as well as for easy maintenance of no adjustment relay. (In the long-haul system, the back contacts are required to prevent simultaneous start of exchange switches in case of circuit fault.)

5. Carrier and signaling-frequency supply system. Designed for small stations, the carrier- and signaling-frequency supply system has capacity to supply eight systems, and is mounted, together with four systems of 12-channel terminal equipment, on one bay. All parts of the supply system which affect all eight systems are provided with normal-emergency automatic switchover function. The channel-carrier supply circuit which affects fewer channels (eight channels) is not provided for emergency use. This is quite desirable for the economical 12×1 direct modulation system.

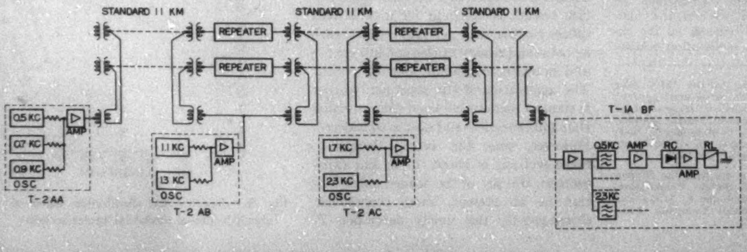


Fig. 5. Diagram of supervisory system

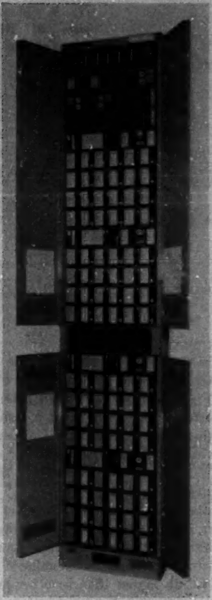


Fig. 6. T-12S-Tr terminal equipped for 48 channels, including carrier supply and termination circuits

6. Complete transistorization and corresponding new assembly system. Employment of transistors, adoption of miniature components, and use of the new type *B* (book) assembly system render the entire equipment quite small and reduce power requirements.

System Design

Modulation System

This system, as discussed previously, allows higher permissible noise than the long-haul carrier system, and consequently permits the allowable inverted crosstalk noise, due to the unwanted side-band frequency introduced by the channel filter, to be comparatively high. Thus, 12 telephone channels are directly arranged in the low-group band, ranging from 12 to 60 kc, by using the same quality *L* and *C* elements as are used in the channel filter NTT has specified for other systems. With this modulation system, the noise allocated to the unwanted side-band frequency (inverted crosstalk noise) is 500 pw for modulation and 500 pw for demodulation. In NTT's pregroup modulation system, 100 pw are allocated to inverted crosstalk noise.

Over-all Noise

The NTT has established a reference standard-toll circuit network, distributing the noise to each section of the reference circuit within technically feasible limits. If the standard for the long-distance transmission system recommended by CCITT is applied to sections between PO (primary outlet) and PO, an over-all noise of 8,000 pw can be allocated to short-haul carrier sections between TC (toll center) and PO. This value ensures sufficiently good speech quality for various kinds of toll networks. The *T-12S-Tr* system has been designed to meet an over-all noise limit of 8,000 pw (at 0-db reference point) for a maximum circuit-length section of 100 km. This over-all noise is divided into two categories, one originating in the terminal equipment and the other in the repeater transmission route.

Line Crosstalk

The short-haul carrier system is designed for an intelligible crosstalk limit of more than SX (signal-to-crosstalk) = 52 db for 100% of the circuits, considering its applicable sections. Intelligible crosstalk is generally composed of far-end crosstalk, loop crosstalk via voice pairs, and intraoffice channel crosstalk. Since intraoffice channel crosstalk is negligible, compared with the other two types of crosstalk, the S/X limit of 52 db+3db=55 db is distributed to each far-end and loop crosstalk. If the limit for far-end crosstalk is not satisfied, selection of pairs or crosstalk compensation will be applied. For loop crosstalk via voice pairs, a suppression coil should, if necessary, be inserted into the circuits to meet the 55-db limit.

Far-end crosstalk not only gives intelligible crosstalk to the circuits, but if the number of carrier pairs in the same cable increases, it will develop into "bubble noise" (so-called multicrosstalk noise) which causes deterioration in circuit quality. Considering the future demand for circuits between PO and TC or TC and TC, it is assumed that the number of carrier systems per route will be about 24.

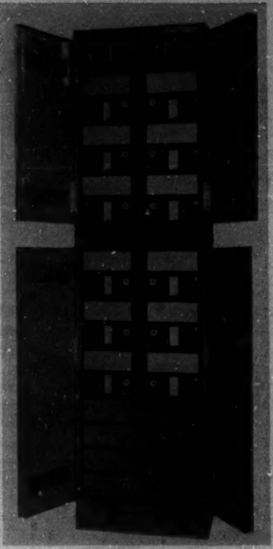


Fig. 7. T-12S-Tr repeater equipped for six systems

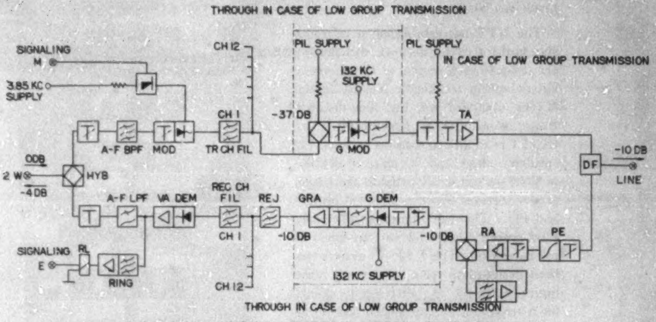
If the line conforms to the limit of intelligible crosstalk, the multicrosstalk noise will be induced by the other 23 systems in the same cable, corresponding to crosstalk distribution. The calculated values of multicrosstalk noise for a circuit length of 100 km are shown in Fig. 4.

Pulsive Noise

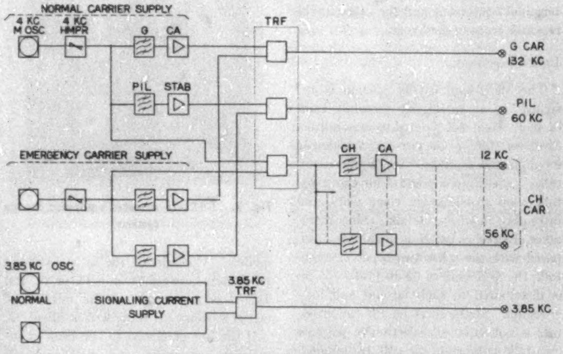
The pulsive noise is peculiar to the short-haul carrier system. It is caused by the harmonic components of d-c dial-

Fig. 8. Channel modulator and demodulator board, including voice amplifier

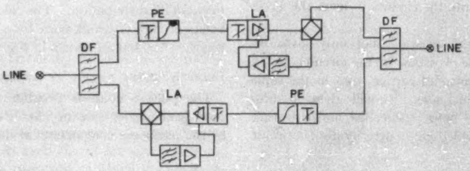
Fig. 9. Block diagram of T-12S-Tr system


THROUGH IN CASE OF LOW GROUP TRANSMISSION
SIGNALING M
3.85 KC SUPPLY
CH 12
PIL SUPPLY
132 KC SUPPLY
PIL SUPPLY
IN CASE OF LOW GROUP TRANSMISSION
CH 1
-37 DB
TA
A-F BPF
MOD
TR CH FIL
G MOD
-10 DB
DF
0DB
2 W
-4 DB
HYB
LINE
REC CH FIL
A-F LPF
VA DEM
REJ
GRA
G DEM
CH 1
-10 DB
-10 DB
RA
PE
SIGNALING E
RL
RING
132 KC SUPPLY
CH 12
THROUGH IN CASE OF LOW GROUP TRANSMISSION

(A) BLOCK DIAGRAM OF TERMINAL EQUIPMENT


NORMAL CARRIER SUPPLY
4 KC M OSC
4 KC HMPR
G
CA
TRF
G CAR 132 KC
PIL
STAB
PIL 60 KC
EMERGENCY CARRIER SUPPLY
CH
CA
12 KC
CH CAR
56 KC
3.85 KC OSC
NORMAL
SIGNALING CURRENT SUPPLY
3.85 KC TRF
3.85 KC

(B) BLOCK DIAGRAM OF CARRIER SUPPLY CIRCUIT


PE
LA
DF
DF
LINE
LINE
LA
PE

(C) BLOCK DIAGRAM OF REPEATER EQUIPMENT

ng, a-c dialing, d-c telegraph current, etc., leaking to the carrier pairs from the voice pairs in the same cable. The pulsive noise is given in the following formula:

$$(S/N)p = -P - L - 10 \log n + S_0 \text{(db)} \qquad (1)$$

where

$(S/N)p$ = signal to pulsive noise ratio which occurs on a 100-km circuit
L = line loss of repeating span
n = number of repeating spans
S_0 = transmission level
P = pulsive noise at the end of one repeating span (per channel)

By actual measurement at the end of the line, a pulsive noise value, P, of about −100 to −150 dbm (power in db referred to 1 milliwatt) was obtained for the 4-kc band at about 120 kc.

Over-all Design

To decide on the permissible over-all noise level of a circuit, a calculation of the component noises should be made. In this system, the repeater span, transmission level, and various other ratings have been determined, so that the over-all noise may meet the most suitable value of 8,000 pw, with a level variation of ±3 db. Fig. 4 shows the over-all noise and the various kinds of noise at a standard repeating span of 11 km. Considering the over-all noise and the repeater overload, −10 dbm has been selected as the transmission level.

As the signaling-current level of this system, −20 dbm has been selected, considering overload during signaling and distortion noises. This value of −20 dbm is 5 db higher than in other multichannel systems. It has, therefore, simplified signaling circuits and contributes to economization of the system.

Supervisory System

This system has been designed to utilize unattended repeater stations. Local power-receiving and floating systems have been adopted by means of rectifiers and storage batteries. The batteries are designed to supply continuous power to the repeaters for 24 hours in case of commercial power failure. Since the repeater station is unattended, alarm equipment for indicating blown fuses, high-low AGC (automatic gain control) alarms, etc., is needed in the supervisory station. The system is provided with such equipment and a service-talking telephone channel.

Ordinary alarm equipment, which transmits information by means of a relay circuit, is not used because it consumes too much power. Instead of a relay circuit, another alarm circuit, which is described below, is used. As illustrated in Fig. 5, the oscillating frequency is continuously generated by a transistor-oscillator as "information source," and in the attended supervisory station a transistor-amplifier, filter, and relays are installed for supervision of the information from the unattended stations. This alarm circuit is composed of a phantom circuit

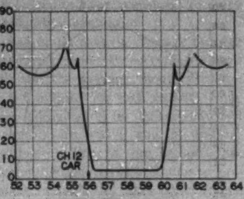


Fig. 10. Frequency characteristics of channel filter

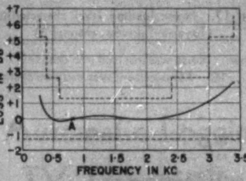


Fig. 11. Typical frequency response of a channel (solid line) and 3/5 of frequency-response limits recommended by CCITT for a single channel of the reference circuit (broken line)

Note: Point A indicates the reference point of 800 cps

from a transistorized short-haul carrier system.

Equipment Design

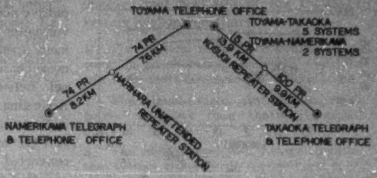
Terminal equipment for four systems is mounted on one standard bay [2.75 m (meters) high]. The bay may also be equipped with the carrier and signaling-frequency supply circuits having emergency sections if needed. Repeater equipment for eight systems can be mounted on one bay (2 m high), including AGC equipment.

The equipment is mounted in type *B* assembly system which has been newly developed by NEC in conformity with transistorization of the carrier equipment. That is, the equipment is assembled on print-wiring boards, each of which is directly plugged into the bay. Figs. 6–8 show the terminal equipment, repeater equipment, and one example of the board (channel modulator-demodulator board).

As described before, the terminal equipment is mounted in the small-station-type mounting system, so-called because used as short-haul carrier equipment, and all necessary parts for the carrier telephone circuit are contained in one bay as shown in Fig. 6.

From the block diagram shown in Fig. 9 and the frequency allocation diagram in Fig. 3, the terminal equipment composition may be understood. Some characteristic parts of it are described in the following.

Fig. 12. Test routes



Audio-frequency filters on the channel-modulator input and voice-frequency amplifier output sides are provided to eliminate the speech-frequency components which would interfere with the ringer, and to prevent signaling-current leakage into the terminating circuit. Channel filters for channel modulation-demodulation have a passband with a frequency range of 300–3,850 cps. These filters are classified into 12 kinds according to the corresponding channel carrier frequencies, and used in parallel connection. Typical characteristics of such a filter are shown in Fig. 10. Typical frequency response of a channel is shown in Fig. 11.

All the amplifiers used in this equipment employ p-n-p alloy-junction germanium transistors. The transistor used for high-frequency amplification has an f_{ab} of 8 mc (megacycles) and Pc_{max} (maximum collector power dissipation) of 140 milliwatts approximately. The others used for low-frequency amplification are classified into two kinds, one of small and the other of medium output.

As an equalization system, a pre-equalizer (4-terminal equalizer) and a 3-stage grounded-emitter feedback amplifier, which has a slope-gain characteristic, are used.

The AGC circuit provides flat regulation to compensate for line-loss variation with weather in the range of ±5 db. It is of the pilot channel system using an indirectly heated thermistor. The highest frequency (60 kc) for low-group transmission and the lowest frequency (72 kc) for high-group transmission are used as pilots, respectively. The mean gain of the repeater is 40 db at the highest line frequency (120 kc), and the maximum gain is 47 db.

The master oscillator used is a 4-kc crystal oscillator which is housed in a small thermostatic oven for frequency stabilization. The distribution amplifier for each carrier current, signaling current, and pilot current, is a bridge-type amplitude-limiting selective amplifier which is effective for suppression of output-level variation and also for improvement of selective characteristics, and hence simplification of carrier band-pass filters.

Commercial Test Data

The NTT conducted a test of the system on the two routes, Toyama–Takaoka and Toyama–Namerikawa, in July 1958; these routes are shown in Fig. 12.

In the Toyama station, terminal equipment for high-group sending and low-group receiving, and in the Takaoka and Namerikawa stations, terminal equipment for low-group sending and high-group receiving was installed. The new-type 2-meter-square unattended repeater station (capable of being equipped with a maximum of 24 systems plus alarm equipment) especially designed for this system, was used for the Harihara repeater station.

EQUALIZATION

Since the maximum repeating span of the system is 12 km, the equalization characteristics test was conducted mainly on the Toyama–Kosugi section (13.9 km). Figs. 13 and 14 show the equalization characteristics between terminal stations, Toyama–Takaoka. In the illustrations, the continuous line shows the character-

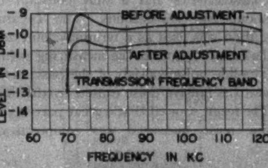


Fig. 13. Equalization characteristics between terminal stations, Toyama–Takaoka

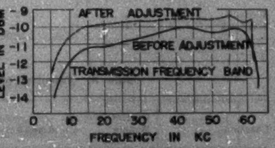


Fig. 14. Equalization characteristics between terminal stations, Takaoka–Toyama

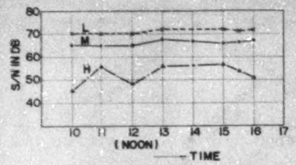


Fig. 15. Line noise characteristics (Toyama–Takaoka, channel 6)

Note: Indication of the noise meter was continuously observed during 1 minute every 1 hour. Three charts show the lowest, L, the mean, M, and the highest, H, pointing value during each of 1-minute observations, respectively

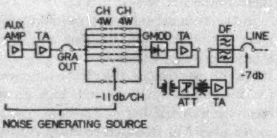


Fig. 16. Measuring circuit for noise loading test

istics equalized by kilometer-step, and the dotted line shows the characteristics obtained after fine-adjusting. In the case of a single 14-km span, the equalization, as shown, will be achieved by fine adjustment.

NOISE

The line noise on the Toyama–Takaoka circuit was measured with a noise voltmeter, which had psophometric weighting characteristics and a time constant of about 250 milliseconds. The Toyama station was used as a terminal point and Takaoka station as the measuring point. The line noise consists of the basic noise and the pulsive noise caused by the line. The result showed that a great part of the line noise measured was composed of pulsive noise, and the pointer of the noise voltmeter vibrated during the measurement operation. Fig. 15 shows the line-noise characteristics measured on a channel between Toyama and Takaoka. According to Fig. 15, the mean value of S/N during a busy hour is about 65 db and the worst value is about 45 db. If converted into the line terminal value by equation 1, the measured mean value will be −110 dbm/channel during a busy hour.

In order to study the over-all noise loading a test was conducted by the measuring circuit shown in Fig. 16. The result of the measurement is shown in Figs. 17 and 18. Measurement was made under the following conditions:

1. To remove the inverted crosstalk noise, the loading noise was loaded on nine channels, excluding the channel to be measured and its adjacent channels.
2. The standard loading noise of $(-1 + 4 \log n = +3.2 \text{ db})$ at the transmitting amplifier out (TA OUT) was inserted. This formula $(-1 + 4 \log n = +3.2 \text{ db})$ shows the standard value recommended by CCITT as an equivalent value for small-channel capacity systems (below 240 channels).

As indicated in Fig. 17, the transmission level of − 10 dbm is a proper value for high-group transmission. As shown in Fig. 18, the basic noise of low-group transmission is less than that of high-group transmission, thus the optimum transmission level for the low group is lower than that for the high group.

The dotted line in Fig. 18 shows the characteristics when the 3.85-kc signaling frequency was loaded on six channels with noise being loaded on the other channels. In this case, almost the same noise value is given as that obtained when the channels were loaded only by noise.

Fig. 17 can be diagramatically analyzed for the basic noise, the third-harmonic noise and the terminal linear crosstalk noise. The basic noise can also be converted into the value for a standard repeating span of 11 km. With this procedure is obtained the signal-to-noise ratio for the 100-km reference circuit, which in this case, will be about 300 pw at the standard level. Even if other noises considered as over-all noise, such as terminal-station noise (inversion crosstalk), multicrosstalk noise, and pulsive noise, are added, the noise-loading level will be about 3,500 pw. Therefore, even with the transmission-channel level variation of 3 db this system conforms satisfactorily to the circuit over-all noise standard of 8,000 pw.

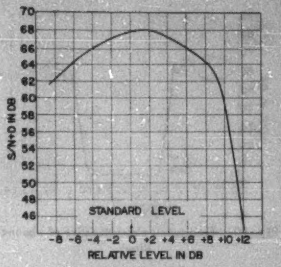


Fig. 17. Noise loading level characteristics (Toyama–Takaoka, channel 3)

Note: Noise was inserted into channels 1, 5, 6, . . . 12. Measurement was made on channel 3

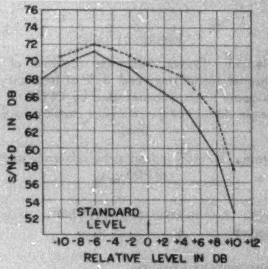


Fig. 18. Noise loading level characteristics (Takaoka–Toyama, channel 1)

Note: Basic noise S/N = 78 db. Rigid line indicates the case of channel 3∼12 noise loading. Broken line indicates the case of channel 3∼6 noise loading and channel 7∼12 signaling current loading

Table I. Number of Failures in Terms of Components

Location	Transistor	Reed Relay	Condenser	Others	Total
Toyama	1	2	10	11	24
Takaoka	4	2	5	5	16

Table II. Number of Failures in Terms of Units

Location	Carrier Supply	Amplifier	Modulator and Demodulator	Channel Parts	Total
Toyama	4	2	8	23	37
Takaoka	1	6	0	18	25

System Reliability

Reliability of the system has been carefully examined during the last 2 years. It was excellent. The 120 circuits on the routes mentioned previously which are composed of 62 toll-dial circuits and 58 nondial circuits have been used to carry out reliability evaluation.

It was found that the average number of failures per month is 0.13 for each circuit and the average monthly failure time is 6 minutes for each circuit, while corresponding figures for our conventional toll circuits are 0.64% and 38 minutes respectively. Thus it can be seen that the transistorized carrier system is several times as reliable as the conventional system.

Table I shows the number of failures in terms of components and Table II in terms of units. As shown in Table I, the number of transistor failures is fairly small and mentioned figures correspond to the failure rate of 0.013% per 1,000 hours. (Total number of transistors used for 120 telephone circuits is about 1,600.) This failure rate may be considered fairly good for transistors in the first stage of development.

Conclusions

The system described in this paper has been operating successfully for more than 2 years, and at present a considerable number of these systems are being put into operation throughout Japan. Performance met the design objectives in a highly satisfactory manner.

Experience gained during factory production and field use of the *T-12S-Tr* system has proved that:

1. The system permits economical constitution of short-haul networks even for relatively short distances (25–100 km).
2. The transistorized system has much better reliability than the conventional one.
3. The newly developed 1-stage modulation system, using ferrite-core coil and polystyrene condenser filters, is fairly effective in cost reduction.

Reference

1. Transistorized Short-haul Carrier System—T-12S-Tr System, O. Ishihara, S. Fujita. *Japan Telecommunications Review*, Tokyo, Japan, vol. 1, Autumn 1959, pp. 32–40.

Automatic Logging of AM, FM, and TV Transmitting Station Parameters

GUSTAVE EHRENBERG
ASSOCIATE MEMBER AIEE

A SYSTEM is described in this paper which automatically and continuously measures and records parameters associated with amplitude modulated (AM), frequency-modulated (FM), and television (TV) transmitters, either remote or local. The technical fundamentals used for recording these parameters are also discussed.

The periodic logging of data as required by the Federal Communications Commission is a routine operation which is best performed mechanically rather than by a man. Not only is a mechanism usually more accurate and more reliable for such purposes, but, in addition, automatic logging of parameters allows the trained, reasoning engineer to devote more time to those jobs which can utilize his skill and judgment to better effect.

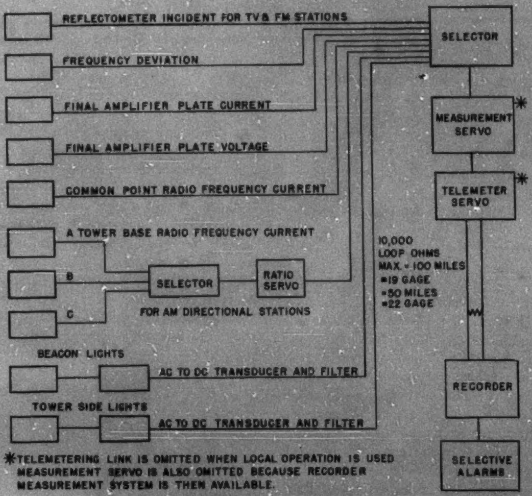
Station policy and other intangibles may weigh significantly in management's decision as to whether to install automatic logging equipment. These factors differ for each station, and are beyond the scope of this paper. However, general industry has found it most profitable to use such mechanisms for typically routine, repetitive tasks, and the broadcasting industry should find this true also.

The following are the transmitting parameters which are usually recorded:

1. Final amplifier plate voltage.
2. Final amplifier plate current.
3. Common-point radio-frequency current.
4. Amplitude ratio of tower-base radio-frequency currents (for pattern indication of directional array).
6. Frequency deviation.
7. Reflectometer incident.
8. Tower beacons and navigation lamps.

Several different packages of equipment are available commercially for this purpose, details of which are given in the Appendix. These are used for:

1. Remote AM stations with directional array.
2. Consolidated AM stations with directional array.


REFLECTOMETER INCIDENT FOR TV & FM STATIONS
FREQUENCY DEVIATION
FINAL AMPLIFIER PLATE CURRENT
FINAL AMPLIFIER PLATE VOLTAGE
COMMON POINT RADIO FREQUENCY CURRENT
A TOWER BASE RADIO FREQUENCY CURRENT
B
C
SELECTOR
RATIO SERVO
FOR AM DIRECTIONAL STATIONS
BEACON LIGHTS
AC TO DC TRANSDUCER AND FILTER
TOWER SIDE LIGHTS
AC TO DC TRANSDUCER AND FILTER
SELECTOR
MEASUREMENT SERVO *
TELEMETER SERVO *
10,000 LOOP OHMS MAX. = 100 MILES #19 GAGE =50 MILES #22 GAGE
RECORDER
SELECTIVE ALARMS

*TELEMETERING LINK IS OMITTED WHEN LOCAL OPERATION IS USED MEASUREMENT SERVO IS ALSO OMITTED BECAUSE RECORDER MEASUREMENT SYSTEM IS THEN AVAILABLE.

Fig. 1. Diagram of the automatic logging system

Paper 60-60, recommended by the AIEE Broadcasting Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted November 2, 1959; made available for printing January 10, 1961.

Gustave Ehrenberg is with the Minneapolis-Honeywell Regulator Company, Philadelphia, Pa.

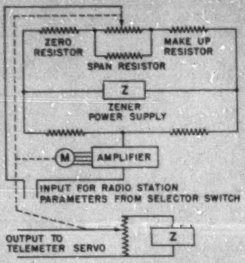

ZERO RESISTOR
MAKE UP RESISTOR
SPAN RESISTOR
Z
ZENER POWER SUPPLY
M
AMPLIFIER
INPUT FOR RADIO STATION PARAMETERS FROM SELECTOR SWITCH
OUTPUT TO TELEMETER SERVO
Z

Fig. 2. Measuring servomechanism

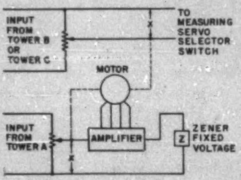

INPUT FROM TOWER B OR TOWER C
TO MEASURING SERVO SELECTOR SWITCH
X
MOTOR
INPUT FROM TOWER A
AMPLIFIER
X
Z
ZENER FIXED VOLTAGE

Fig. 3. Ratio servomechanism

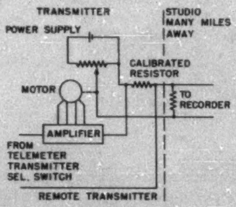


Fig. 4. Telemeter servomechanism

3. Remote AM, FM, or TV stations (omnidirectional).
4. Consolidated AM, FM, or TV stations (omnidirectional).
5. Minimum systems (similar to type 4).

Technical Description of the System

In the automatic logging system (Fig. 1), voltages from the various parameters, proportional to their values, are converted into low direct voltages or direct millivoltages, and fed to the selector switch. Each voltage or millivoltage, in sequence, then feeds through the switch and on to the measuring servomechanism. From there, the signal is placed on the telemetering servo which transmits it a maximum of either 100 miles over a no. *19* Awg (American wire gage) telephone pair or 50 miles over a no. *22* telephone pair. Since the signal is direct current, the least expensive type of continuous metallic pair is adequate; however, maximum resistance cannot exceed 10,000 loop ohms.

The transmitted signal of 1.5 to 3 milliamperes is changed into millivolts by a dropping resistor at the receiver. The receiving recorder then takes each millivolt signal in proper sequence and prints it on the chart, each with its own distinctive identifying color and symbol. Once each revolution the selector switch in the receiver is synchronized with the selector switch in the transmitting unit.

There are eight selective high or low alarm points which can be wired to any recorder point at the customer's discretion. Two alarm points can be used to bracket a parameter reading with a high and a low alarm. As many as 15 different variables (parameters) can be recorded in sequence in multicolor on the 11-inch-wide strip chart at the rate of one point each 30 seconds.

The automatic logging system will accept direct voltage in any of the following standard ranges: 0–10 millivolts, 0–1 volt, 0–2 volts, 0–5 volts, 0–8 volts, 0–10 volts. If zero-center scales are desired or required, these voltage ranges can be suppressed.

Technical Description of Components

TRANSMITTING UNIT SEROVMECHANISMS

The transmitting unit measuring servo (Fig. 2) uses a bridge circuit to supply standard variable millivoltages to null-balance the incoming signal. The bridge is powered by a Zener diode voltage supply which is accurate to 0.1%. A retransmitting slidewire, also powered by a Zener supply, gives an electric signal to the telemetering servo input. The servo is accurate to 1/4%.

The transmitting unit ratio servo (Fig. 3) accepts two voltages: one from tower A and the second from tower B or C in sequence. The tower A voltage is servo-balanced against a Zener voltage, Z. Tower B voltage is brought to the retransmitting slidewire which is mechanically ganged to the first slidewire. Thus

$$XA = Z$$

$$X = \frac{Z}{A}$$

$$XB = \text{output voltage}$$

Substituting Z/A for X,

$$\text{Output voltage} = \frac{Z}{A} B = \frac{B}{A} Z$$

As Z is a constant and can be disregarded, the servo position is proportional to B/A.

The output voltage is delivered via the selector switch to the measuring servomechanism with an accuracy of 1/4%.

The transmitting unit telemeter servo (Fig 4.) transmits a milliamperage proportional to an incoming millivoltage. A null balance must be achieved between the incoming millivoltage and the outgoing milliamperage passing through a calibrated fixed resistor in series with the transmission line. The milliamperes through the resistor produce millivolts. When the system is not in balance, the servo-amplifier drives the rheostat on the power supply to increase or decrease the voltage until the proper milliamperage is transmitted to give a balance. Obviously, this makes the telemetering system entirely independent of power-supply voltage fluctuations, and also entirely independent of transmission-line resistance change as long as it does not exceed 10,000 ohms. Line leakage is the only effect not compensated for in this type of telemetering system. On a good line, the system is accurate to 1/4%.

SERVO-OPERATED RECORDER FOR RECEIVING UNIT

In the servo-operated recorder (Fig. 5), a synchronous vibrator is used to "chop" direct current into pulses going through an input transformer that is center-tapped to produce alternating current. The alternating current is applied to the grid of the first voltage-amplification stage. Note that this alternating current is in phase with the 115-volt line supplying current to the transformer of the power amplifier. The synchronous vibrator has a small permanent magnet operated by an electromagnetic coil which is energized by the power-transformer filament winding.

The amplifier consists of three standard resistance–capacitance-coupled voltage-amplification stages. The output stage is actually a phase discriminator. The grids of the two sections of the output tube are connected together and are operated from the output of the last voltage stage. The plates of the tube are supplied with alternating power. The cathodes of the tubes are connected to the end of one winding of a 2-phase motor, the other end of which connects to the center tap of the transformer supplying the plate voltage. The second winding of the 2-phase motor is energized from the 115-volt line supplying power to the amplifier. This winding is in series with a capacitor which shifts the phase of the current by 90 degrees. This phase shift is needed for operation of a 2-phase motor. One field should be 90 electrical degrees from the other. The phase of the current coming from the amplifier will be either in phase with the power-line current, or 180 degrees out of phase because the synchronous vibrator remains in phase with the lower line.

1. If there is no voltage input to the amplifier, the grids of the output tubes will not be energized and the output tubes will pass current, since they are not at cutoff. However, this will be 120-cycle direct current as shown. The servomotor will then tend to move in opposite directions for each half-cycle, but will not move.

2. If the input to the amplifier is plus-minus, the grids of the output stages will be plus when one tube fires and minus when the other tube fires. Therefore, the output will appear similar to the second waveform (Fig. 5). Since one set of waves has grown and the other has diminished, the motor will move in the direction designated by the large waves.

3. If the input to the amplifier is minus-plus, the grids of the final output stage will be positive (+) when the opposite tube plate is positive or when the other half-cycle appears. Therefore, the output wave will be similar to the third waveform, or 180 degrees from the first set of half-cycles. The motor will turn in the opposite direction.

A capacitor in parallel with the winding of the motor energized by the amplifier tunes the circuit to 60 cycles. Instead of a half-wave pulse, the oscillations produce a full 60-cycle alternating-current sine wave which can be seen on an oscilloscope. The capacitor, in series with the other motor winding energized by the 115-volt power line, tunes the winding for minimum impedance. This winding draws approximately 11 to 15 watts from the power line while the winding energized from the amplifier draws approximately 1 to 2 watts. This is sufficient to supply a rotating field to the motor. The amplifier has an amplification factor of 1,000,000. About 4 microvolts across the input will start the servomotor. The recorder pen, as well as a series of adjustable cams which trip switches for alarms, are operated by this servo system.

Alarm System in Recorder

The alarm system is a selective type, so that only the alarm connected with the point being measured can operate. This allows each variable to operate anywhere on the instrument's 11-inch scale without affecting any other point. The cam alarm switches are energized in sequence through a selector switch. A timing mechanism allows the instrument to come to balance before applying power to this circuit. Relays are dropped out by the cam switches to give an alarm light if the cam settings are exceeded. A manual reset button picks up the relays, which are then held in by a holding contact.

Transducers

A number of transducers are required at the transmitter to give signals to the logging equipment. One of these, the tower-light transducer (Fig. 6), is a current transformer whose primary is placed in series with one line of the tower-light cable. The output is across a variable resistor producing an alternating voltage which is rectified by a full-wave bridge before it passes through two choke capacitor filters and a radio-frequency filter. The smoothing filter will take a 40-cycle per-minute pulsing of tower beacons and give only a 1% flicker to the pen motion of the recorder. The same transducer is used for tower side lights and beacon lights.

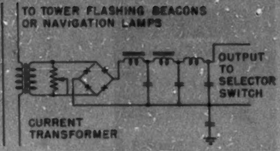

TO TOWER FLASHING BEACONS OR NAVIGATION LAMPS
OUTPUT TO SELECTOR SWITCH
CURRENT TRANSFORMER

Fig. 6. Tower-light transducer

Voltage transducers are generally located in the transmitters, especially if remote control is used. They are voltage dividers to reduce voltage to a low value at ground reference. Current transducers are shunts, usually in the ground side of circuits, and of sufficient size to give a voltage if remote control is used.

For large current measurement at high voltage above ground reference, a saturable reactor, powered by the 120-volt alternating-voltage power line, isolates the high-potential current electrically. The reactor requires a full-wave rectifier and filters, plus voltage division, to reduce the input to an acceptable level for the logging equipment.

The monitors for frequency, reflectometer incident, etc., include a meter circuit for remote reading. Generally, provision is made to substitute an external meter for a 10% resistor. However, a calibrated resistor in the range of the external meter resistance can be nserted instead to given an adequate millivoltage for either microamperes or milliamperes in the meter circuit. The logging equipment uses this millivoltage directly. This millivoltage is generally at ground reference, and where it exceeds 100 volts, an isolation servomechanism with ground reference retransmission is required.

Summary

The automatic logging of AM, FM, and TV transmitting stations is now an accomplished fact. Such systems are presently operating in stations *WTOP* and *WMAL* in Washington, D.C., and *KFI* in Los Angeles, Calif.

Data gathered from the *WTOP* and *KFI* installations have been submitted by the National Association of Broadcasters to the Federal Communications Com-

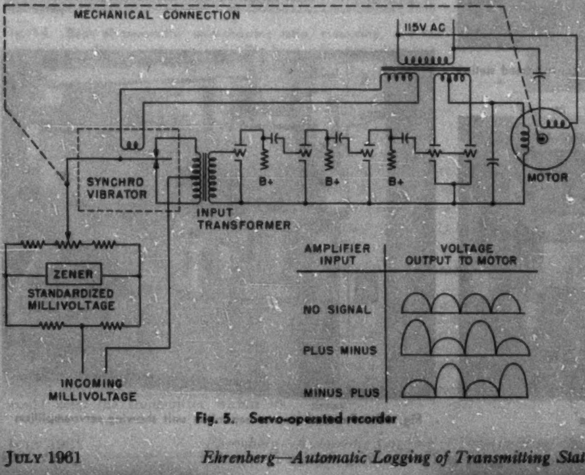


Fig. 5. Servo-operated recorder

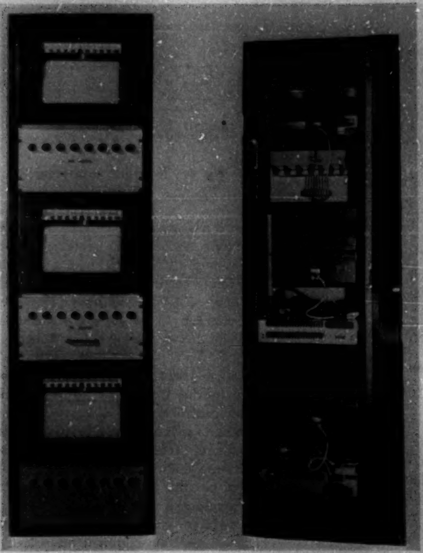


Fig. 7. Receiving instruments with alarm light panels

Fig. 8. Light panel with input adjustments for a consolidated station

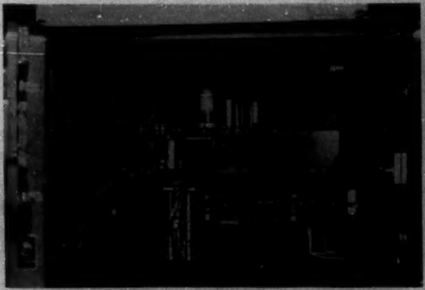


Fig. 9. Inside of recorder case

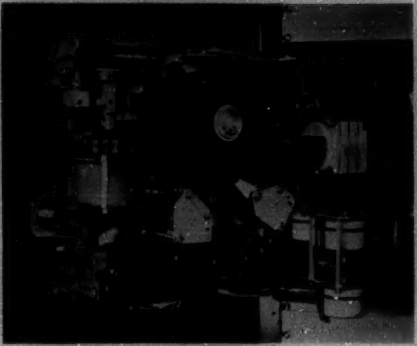


Fig. 10. Receiving recorder chassis swung out showing selective-alarm adjustable alarm switches

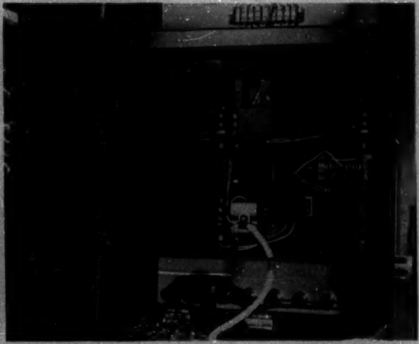


Fig. 11. Alarm relays on back of receiving recorder

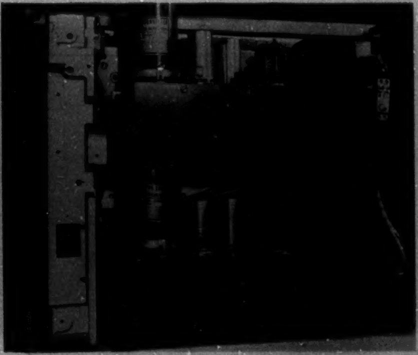


Fig. 12. Front of remote transmitter unit showing servoamplifiers

Fig. 13. Transmitter showing front panel swung out. Note input selector switches. Space is available for extra switches for automatic control

Fig. 15. Side view of transmitter showing amplifiers, servos, and input adjustments

Fig. 14. Back of transmitter unit showing ratio, measuring, and telemeter servos, input adjustments, navigation-lamp calibrating equipment, and input terminal blocks

Fig. 16. Side view of transmitter showing sequencing gears

mission to secure a ruling that these charts are acceptable as the official log of a station. What has long been suspected is being proved. Improved accuracy and reliability of readings, as well as improved efficiency in the use of manpower, is making the use of automatic logging equipment mandatory in station operation.

The system and its component parts have all been designed with an eye toward the future. Provisions have been made in the remote measuring and telemetering unit for the installation of automatic control for transmitters. This control would keep the transmitter within its parameters right at the transmitter site, and so would require no long-distance telemetering for control purposes. This system is available now, if the broadcasting industry will specify it.

Appendix. Commercial Packages

Equipment packages are available commercially for several different applications; all are designed for relay rack mounting. The components of the various packages are detailed in the following.

1. Remote AM station with directional array:

A. Transmitter assembly, 20-inch-high vertical panel space.

- (1) Measurement servo.
- (2) Ratio servo.
- (3) Telemeter servo.
- (4) Synchronizing apparatus.
- (5) Radio-frequency filters.
- (6) Input circuit span calibrating potentiometers.

B. Receiver; $15^{3}/_{4}$ inches high, vertical.

- (1) Strip chart recorder with multipoint print-out and selective alarm switches.
- (2) Eight-alarm lamp panel; $3^{1}/_{2}$ inches high, vertical.

2. Consolidated AM station with directional array:

A. Multipoint strip chart recorder with measuring circuit and selective alarm switches.

B. Eight-alarm light panel; $3^{1}/_{2}$ inches high, vertical.

C. A $10^{1}/_{2}$-inch-high vertical panel containing:

  (1) Ratio servo.
  (2) Radio-frequency filters.
  (3) Input circuit span calibration potentiometers.

3. Remote AM, FM, or TV station (omnidirectional):

A. Transmitter assembly; 20-inch-high vertical panel space.

  (1) Measurement servo.
  (2) Telemeter servo.
  (3) Synchronizing apparatus.
  (4) Radio-frequency filters.
  (5) Input circuit span calibration potentiometers.

B. Receiver.

  (1) Multipoint strip chart recorder with selective alarm switches.
  (2) Eight-alarm lamp panel; $3^{1}/_{2}$ inches high, vertical.

4. Consolidated AM, FM, or TV station omnidirectional):

A. Multipoint strip chart recorder with measuring circuit and selective alarm switches.

B. Eight-alarm light panel; $3^{1}/_{2}$ inches high, vertical.

C. A $10^{1}/_{2}$-inch-high vertical panel containing:

  (1) Radio-frequency filters.
  (2) Input circuit span calibrating potentiometers.

5. Minimum system (similar to type 4):

All of the equipment described above includes as standard a series of calibrating precision rheostats so that the customer's parameters can be brought to within the spans of the measuring system by the turn of a dial. In this system the calibrating rheostats are deleted in order to provide a minimum-expense package for the smallest stations. The station engineers must put in their own calibrated voltage-dividing resistors to bring all of their parameters to the required millivolt span of the measuring system.

Systems 1 and 2 would not be used for FM or TV stations because these stations do not have directional antenna arrays.

Automatic Dialing and Electronic Selection on Railroad Message and Dispatcher Circuits

OTTO A. JORGENSEN
NONMEMBER AIEE

Synopsis: Low-level tone-signal techniques are used for automatic calling and selection on multiparty single-line facilities. The subscribers, equipped with transistorized transmitters and receivers, transmit single tone signals for seizure, dialing, and release. These signals are propagated along the wires to the terminal equipment, where they are converted to d-c pulses and stored for interpretation. This interpretation is accomplished by relay logic in order to decide whether the call is for a party on the same line, or for a party within the PBX (Private Branch Exchange) to which the line is terminally connected. In the first case, the called number is converted into an amplitude-modulated tone, where the frequency of the carrier and the modulation correspond to the first and second significant digits dialed, respectively. This tone is then retransmitted for selection of the desired party. At the called subscriber's station, the electronic selector receiver, which takes the place of the electromechanical selector, responds only to the particular carrier corresponding to its number. In this receiver, the envelope is detected and the recovered modulation tone is amplified directly to drive one out of a maximum of five tuned ringers. A total of 60 stations can be used on one pair of wires having a maximum attenuation of 30 db (decibels).

A VITAL PART of the communication for railroads is carried over their message lines, which run adjacent to the railroad tracks. A party-line system is used because the messages are normally of short duration and are sent progressively to, or from, the stations along the line.

Several distinct functions are performed to obtain a connection; see Fig. 1. First, the calling party at the station transmits a signal to indicate to the terminal that a connection is desired; the term "seizure" is applied here, meaning an electrical connection has been established through operation of the first switch. The calling party then sends information as to the destination of his call, whereupon the terminal transmits a signal, to which the selector immediately reacts by actuating a sounding device that alerts the called subscriber. Finally, upon completion of the call, signals are transmitted to restore automatic equipment to normal. When the apparatus used in the connection is disengaged and restored to its condition when not in use, the term applied thereto is called "release."

In equipment which has been in use many years, several of the operations are done by human interaction, and the called station is selected by electromechanical stepping switches. Now, however, the trend towards faster and more reliable service has required a higher degree of automation where electromechanical functions are replaced by electronic operations.

Components of the New System

Calling Equipment

For calling, stations are now equipped with dial transmitters instead of the previously used, manually operated magneto generators.

A dial, a transistorized oscillator, a high-impedance handset, and a slow-release relay are installed in the housing of a conventional telephone. When the subscriber dials, the dial-pulsing contacts, in this case, "make contacts," activate the 4.5-volt battery-operated oscillator, which sends bursts of 2,350-cps (cycles per second) tone corresponding to the dial pulses to the line. The line is temporarily connected to the oscillator through the off-normal contacts on the dial.

The first digit is interpreted at the terminal as a seize signal, which activates the control unit at the terminal for reception and interpretation of further dial tone bursts.

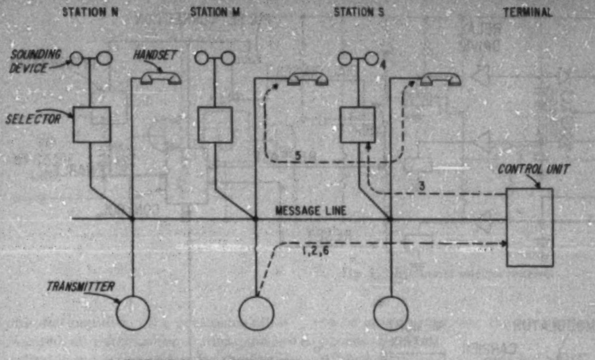
The normal level of the dial transmitter is −5 dbm (power in db referred to 1 milliwatt). This level is sufficient for most practical applications, but for high line attenuation the level can be raised by strapping to +5 dbm, enabling the

---

Paper 61-115, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 28, 1960; made available for printing January 10, 1961.

Otto A. Jorgensen is with Stromberg-Carlson Company, Division of General Dynamics Corporation, Rochester, N. Y.

Acknowledgments are made to M. Zeller of the Northern Pacific Railway Company, who suggested the scheme of using tuned ringers; to N. Butkowitch of the Chicago and North Western Railway, who gave many valuable suggestions; and to R. Staw of the New York Central System, who co-operated in making field tests. B. Gatsert and G. Jowiec of Stromberg-Carlson Company participated in much of the groundwork.



1-SEIZE SIGNAL FROM TRANSMITTER
2-INFORMATION ABOUT DESTINATION OF CALL
3-SELECTOR SIGNAL FROM CONTROL UNIT
4-SOUNDING DEVICE OPERATES
5-CONVERSATION
6-RELEASE SIGNAL TRANSMITTED (WHEN NEEDED)

Fig. 1. Typical message line communications system

subscriber to call over distances up to 250 miles.

At the end of dialing, provision must be made for transmission of a release signal. As the equipment is designed to be operated as nearly as possible according to conventional telephone practices, this tone is produced automatically by a relay arrangement when the telephone is placed on hook; see Fig. 2.

The relay in the telephone instrument has both a high- and low-impedance winding. The latter is energized when the subscriber goes off hook and operates the dial. As soon as the relay is operated, a make contact connects the two windings in series for reduced current drain. Next, the dial-pulsing contacts will periodically connect the battery to the oscillator, which then sends tone bursts to the line via the off-normal contacts. When dial operation is over, the oscillator is nonoperated and disconnected from the line until the phone is replaced on hook. The hook-switch contacts will now re-establish battery to the oscillator and connect the output of the oscillator to the line. This is possible because the relay remains operated for approximately one second due to the capacitor across its winding.

The arrangement of the handset, which in the idle condition is disconnected from the line, also is shown in Fig. 2. The transformer raises the impedance of the handset to about 5 kilohms so that the bridging loss will be low. The press-to-talk key connects the resistor to the receiver for proper sidetone level.

TERMINAL EQUIPMENT

The control unit is designed for operation with either the type *60* electromechanical selectors or the newly developed electronic selectors at the called stations. The discussion here will be concerned mainly with tone selection, but it is pointed out that for type *60* selector operation, the signals are identical to those produced by the present keying machines where $3^1/_2$-cps square waves of high voltage are sent to the station selectors.

The control unit is physically divided into two separate panels working together. One, the common panel, contains tone receivers, pulse correction networks, and stepping switches for counting and encoding the digits stored before transmission of the selector signals. In addition, other relays are used for timing purposes and for connecting the message line to the local PBX. The other panel is the selector panel which, at the user's option, is equipped for either type *60* selector operation or for electronic tone selection.

For type *60* operation, the selector panel contains the switches necessary to produce the $3^1/_2$-cps high-voltage signals. For electronic tone selection, it contains tone oscillators and a modulator for generation of the modulated wave.

DIAL-PULSE RECEIVERS

In the common panel of the control unit, Fig. 3, a narrow band-pass filter allows only 2,350-cps frequency components to enter the dial-pulse receiver. Here they are first amplified and then rectified in two separate channels, one with fast charge and discharge which drives the dial relay and another with slow charge and fast discharge, responding only to steady tones, driving the release relay. To avoid speech components at 2,350 cps from being interpreted as signal tones, a voice-guard receiver amplifies

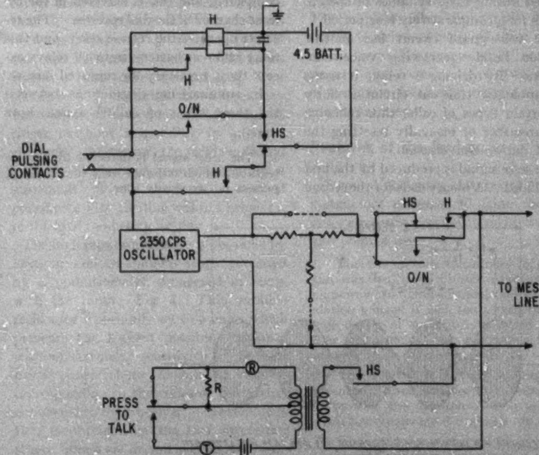


Fig. 2. Station transmitter with automatic hook-switch release and handset

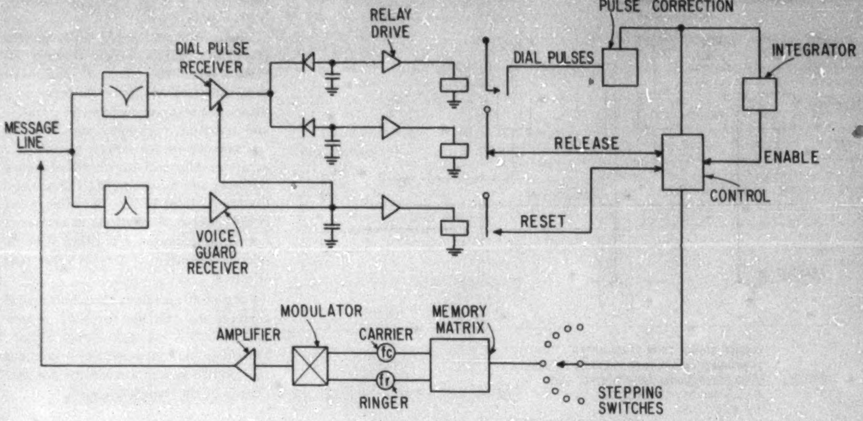


Fig. 3. Control unit

energy of voice frequency other than 2,350 cps to produce an inhibit signal, which disables the dialing receiver when speech is present. The operation of this depends on the fact that the normal voice contains many different frequency components simultaneously.

The inhibit signal is derived from the output of the receiver by a detector with fast charge and slow discharge—20 msec (milliseconds) and 1 sec (second), respectively. In this way "talk downs" are minimized because inhibit signals, developed during early syllables of speech, remain for a comparatively long period.

The voice-guard circuit has another function besides canceling voice simulations. By driving a relay, it resets the automatic time-out circuit in force for certain types of calls, thus relieving the subscriber of manually resetting the circuit during conversation.

The seize signal is produced by the first digit dialed. A binary code for the second and third digits is set up on a relay matrix by the control unit, where an integrating network is preadjusted so that the presence of more than five consecutive pulses from the dial relay will produce an output. This activates the unit, and switches the output of the dial relay so that succeeding pulses, transferred to the rotary switch, will set up the binary code.

The processing of this stored code will be explained after handling of the release signal is understood. First, a steady tone is transmitted as soon as a call is completed, and this is detected in the release channel of the dial receiver. The receiver operates the release relay, and this relay removes battery from all relays except those necessary for repeated seizure.

In summarizing distinctions between the three types of calling signals, note that:

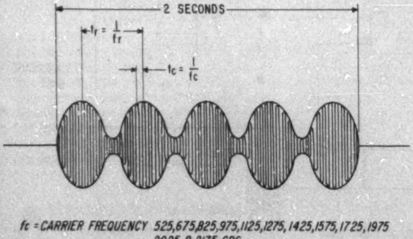
1. The seize signal is detected by the integrator, which responds only to a series of pulses; not to steady tone.
2. The dial pulses are effective only after the control unit has been seized.
3. The release mechanism is effective only upon reception of steady tone.

Selector Unit at Terminal

After the third digit is dialed, the memory matrix sets both carrier and ringer oscillators. The latter has five different frequencies corresponding to a nonharmonic tuned ringer series. The former can be set by the relay matrix for 12 different frequencies, beginning at 525 cps and ranging upward at distances of 150 cps apart. Choice of a linear distribution of frequencies was determined by the fact that the inductive components of the filters involved had quality factors which increased almost linearly with frequency.

The 5 ringer frequencies and the 12 carrier frequencies can be combined in 60 different ways, thus enabling the customer to dial any of 60 stations.

After the dialed code has been read out of the memory unit, a timing circuit turns the oscillators on for 2 seconds. The oscillators feed into a modulator, where the carrier is modulated approximately 80% by the ringer frequency, resulting in a waveshape such as shown in Fig. 4. This wave, transmitted to the message line for 2 seconds at a level of 0 dbm, will be heard by the calling subscriber as a shrill sound, indicating that the desired station is being rung. Immediately after the modulated tone ends, the control unit restores to normal. It is now susceptible to further dialing as long as no conversation is taking place; speech would inhibit



fc = CARRIER FREQUENCY 525,675,825,975,1125,1275,1425,1575,1725,1975 2025 & 2175 CPS

fr = RINGER FREQUENCY 20,30,42,54 & 66 CPS

Fig. 4. Typical modulated wave

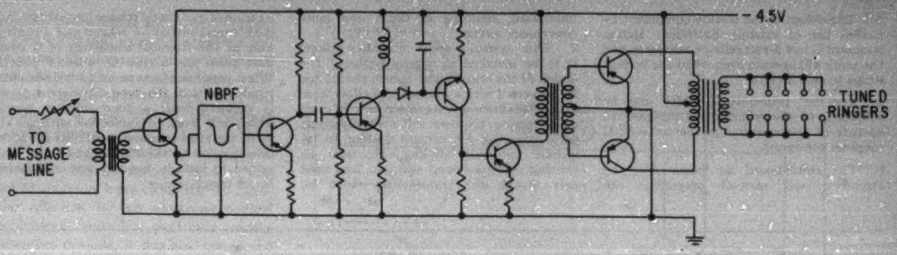


Fig. 5. Electronic selector receiver

the dial amplifier. If a conference call is wanted, all participating stations can be called in succession as long as the line is silent.

The operation just described refers to revertive calls, where the message line can be regarded as a party line.

Connection to PBX

The control unit can also be connected to the local PBX. The message-line subscriber first dials the seizure digit, as for revertive calls, and then dials the PBX access digits, for instance, 09. These digits are detected in the storage matrix which connects the message line to the PBX tip and ring leads. A hybrid coil prevents the PBX dial tone from entering the voice-guard amplifier, where it would produce inhibit signals, barring further dialing. The *E* lead from the PBX trunk circuit is connected to the output of the dial receiver, via the pulse correction network. When dial tone is heard, the subscriber continues to dial in a normal fashion. In this case, the *M* lead from the trunk provides answer supervision.

When the operator is dialed, the answer supervision is absent and the continuity of the connection depends on the control unit, where a 1.5-minute time-out circuit is reset by the voice-guard relay as long as conversation is in progress. This feature is also operative when connections are made to other facilities, such as manual offices or other message lines which do not provide answer supervision.

Use of Repeaters

Some railroad message lines use "one-way acting repeaters" to make up for excessive line loss. This is a condition under which the PBX access is complicated. The dial tone going out from the PBX will set the repeater for high gain towards the calling subscriber on the message line and, therefore, he will not be able to overcome the loss in the opposite direction.

If such facilities are to be used, the PBX trunk circuit must be altered so that dial tone is prevented from entering the message circuit or is limited to a short duration.

Selector Receiver

The selector receiver is a unique design made possible only by the advent of the transistor. Several obstacles had to be overcome to develop a successfully working unit, which, it was found, should be powered from dry cells or rechargeable nickel cadmium batteries with a maximum instantaneous power drain of 1.5 amperes and a stand-by power drain of less than 10 ma (milliamperes). Sufficient power output should be available to drive one of five conventional standard-tuned ringers connected in parallel. The receiver should have enough sensitivity so that it will operate at voltage levels between 0 and −30 db, referred to 0.78-volt rms (root mean square). The bandwidth of the narrow band-pass filter should be wide enough to pass a frequency band of twice the frequency of the highest ringer, without responding to adjacent channels. Moreover, the receiver should have very high input impedance so that a maximum of 60 receivers connected permanently to the line will not present a problem of too-heavy loading.

The high input impedance is obtained by using an emitter-follower stage, preceded by a potentiometer connected in series with the input; Fig. 5. This variable resistance is normally set to a value which presents the highest possible resistance without sacrificing sensitivity. The filter passes only those modulated carriers corresponding to the codes of the parties connected to that receiver. Preamplification is obtained in the two succeeding stages, which are driven in class *A* for constant gain. When a modulated carrier is present, an alternating voltage, corresponding to the envelope frequency, and a superimposed d-c component will appear at the output of the envelope detector. The d-c component is maintained in order to turn on the two succeeding directly coupled stages which, in their idle condition, draw very little stand-by current. The a-c component of the rectified signal is also amplified in these stages, the output of which is sufficient to drive the class *B* push-pull driver stage.

Some distortion is unavoidable in these various stages. This is why a non-harmonic ringer series was chosen so that second and third harmonics of one ringer frequency will not have sufficient amplitude to drive higher frequency ringers. Up to five ringers with frequencies of 20, 30, 42, 54, and 66 cps, respectively, can be coupled to the output. This means that, at locations where several telephones are in the same local area, one receiver can be utilized for as many as five stations. The amplifier also can supply enough energy to drive as many as three ringers of the same frequency when no other ringers are connected, or two ringers of the same frequency when two other ringers are used.

Features of the System

Some of the main features of the new system are summarized from the text, but others, not previously mentioned, are also pointed out.

1. The use of tone techniques for calling and selection has facilitated employment of microwave for transmission, and the modulation scheme is such that, even though a high degree of stability is required for the tone frequencies, no synchronization is needed for the multiplex gear.
2. The system is directly applicable to repeatered lines, and does not require any by-pass of the repeaters, as is necessary when low-frequency d-c signals are used.
3. The system was engineered to conform to normal telephone practices; therefore, minimum instructions are needed when opening new installations.

4. The design of the station equipment included use of existing batteries. Hence, standard 4- to 4.5-volt cells are adequate, and the units will operate even when the voltage drops to near 3 volts.

5. The dial transmitters, which may be subjected to severe cold, are designed to operate at temperatures as low as −40 C (degrees centigrade).

6. The switchboard no longer must be attended, and manual operations are minimized, resulting in faster and more convenient service.

7. This system's great flexibility allows it to be installed in several phases. The design of the control unit is such that it can be converted with a minimum of effort from type *60* selection to tone selection. In cases of conversion, the user will normally begin with extending his present system to include automatic dialing. He uses his existing telephones and selectors, and need only install dial transmitters (with no handset or automatic release mechanism, but with a pushbutton for release and a control unit at the terminal consisting of a common panel and a type *60* selector panel). When expansions are made to include electronic selection, the type *60* selector panel is replaced with a tone selector panel, which will work in connection with the already installed common panel. At the same time (or later), the dial units can be converted to include handset and automatic hooks-switch release.

The British JERC Development of Electronic Switching

T. H. FLOWERS
NONMEMBER AIEE

G. FORSHAW
NONMEMBER AIEE

THE TELEPHONE service in the United Kingdom is operated by the General Post Office and supplied with equipment by a number of manufacturing companies. In general the industry manufactures to specifications issued by the Post Office. In this way a multiplicity of types of apparatus is avoided despite the existence of a number of independent manufacturing concerns offering similar products. In particular, a joint Post Office and industry organization, the British Telephone Technical Development Committee (BTTDC) is responsible for the development and standardization of exchange switching equipment, not only of its technical performance but also very often of its mechanical and electrical design and construction.

Although committed in this way to a common system for manufacture, independent research and development are pursued by all parties with the object of producing new products which will be accepted by the BTTDC for standardization. The application of electronics to telephone switching started in this manner. The fields in which the various parties worked and became knowledgeable included:

1. Semielectronic systems using crossbar switches and high-speed motor uniselectors.
2. Gas-discharge tube space-division exchanges.
3. Control systems using magnetic drums.
4. Computer techniques including storage, logical operations, and programming.
5. 100-channel 4-wire tdm (time division multiplex) exchange systems.
6. 24-channel 2-wire tdm exchange systems.

JERC Organization and Program

The Post Office and its five principal suppliers of exchange equipment, having all worked independently on the fully electronic exchange problem, came to the conclusion that the size of the task required a joint effort if a useful solution was to be had in an adequately short time. After negotiating arrangements whereby the costs and the results of such an enterprise would be fairly shared, in 1956 the Joint Electronic Research Committee (JERC) was established. The main objects are to produce a fully electronic exchange system or systems which will be competitive with existing electromechanical systems, and to explore the effects which electronic operation will have on telephone practice in the future.

The first decision was to put effort not only into research but also into exchange design development. Research covered space-division, 2-wire tdm, and 4-wire tdm systems, with general problems left to all parties to pursue as they were able. Among the parties the 4-wire tdm type of exchange with highway switching was the most advanced in philosophy and technique, and was chosen for the joint exercise in design and construction and, ultimately, operation in service of an electronic exchange to be known as the Mark I (Mk I) tdm system.

The Mk I system has absorbed more effort than was expected and has thereby delayed the more general research work. It is described in this paper in sufficient detail for an understanding of the system and of the conclusions to be drawn from the experiment. The paper also discusses the problems associated with introducing electronic exchanges into existing networks, the present trend of thought concerning space division and Mk II tdm systems, and some observations on the possible effects of electronic exchanges on network planning.

Mk I TDM 100-Channel Switched Highway System

Resulting from the policy decisions made by JERC, a co-ordinated design effort was brought to bear on a working model to prove the practicability of the design, and then on final engineering design for a field trial exchange to be installed for public service at Highgate Wood, in North London, by 1962. Based on experience at that time, decisions had to be made to plan the outline of the system, including a survey of suitable devices to be incorporated into it. This resulted in the adoption of magnetostrictive delay lines as temporary stores and magnetic drums for semipermanent storage. Due to the disposition of the individual labor forces employed on the development, the detailed design was split into suitable sections and the techniques with which the respective parties were most familiar were incorporated in their design. This has resulted in the use of a wide variety of techniques and components, and while it has introduced com-

---

Paper 61-73, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 2, 1960; made available for printing November 30, 1960.

T. H. Flowers is with the British Post Office, London, England, and G. Forshaw is with the Automatic Telephone and Electric Company, Ltd., Liverpool, England.

Acknowledgment is made to the engineer-in-chief of the British Post Office and to the companies involved (Associated Electrical Industries, Ltd.; Automatic Telephone and Electric Company, Ltd.; Ericsson Telephones Ltd., Beeston; Standard Telephones and Cables Ltd.; and The General Electric Company, Ltd.), for permission to make use of the information contained in the paper.

plicated interworking and power supplies, it is expected that valuable experience will be gained in comparing the operation and reliability of the several sections. Primarily, the division of design effort has facilitated a field trial in the shortest possible time.

While the test exchange was to be within the London director network and would therefore have to satisfy the facilities afforded by an electromechanical equivalent and by electronic means wherever possible, it was also considered necessary that the system should be capable of serving up to 20,000 lines without substantial change of technique. Because of the experimental nature of the public service field trial, arrangements are made to provide an electromechanical stand-by with all lines connected via changeover relays. This covers periods when the system might be out of service for modification, and it will finally take its place at the end of the field trail. Precautions have been taken to ensure that subscribers will not know from which equipment they are receiving service.

The choice of Highgate Wood has required a design for a maximum of 2,400 line equipment positions within the expected life of the field trial, i.e., up to the year 1972. It caters to 27 different types of line: ordinary, shared, PBX (private branch exchanges), CCB (coin-collecting boxes), junctions, etc. The problems of compatibility, maintenance, and security of service encountered resulting from the system's being introduced into an existing network have involved a considerable part of the design effort.

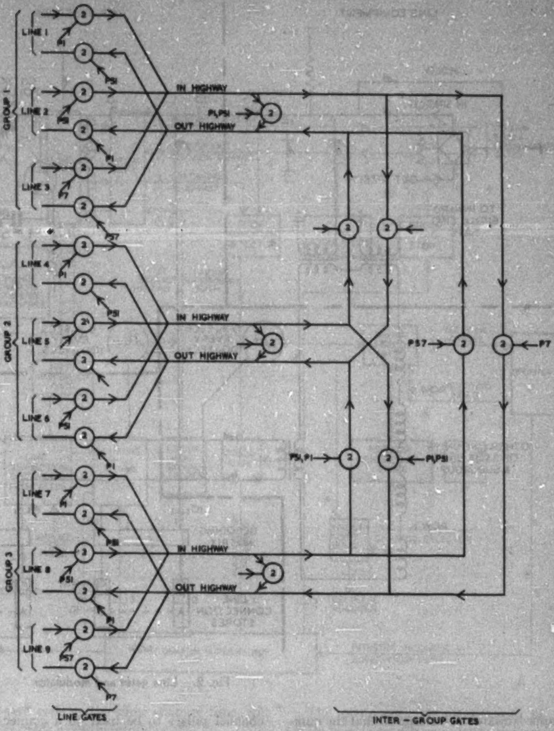


Fig. 1. Basic trunking

OUTLINE OF THE SYSTEM

The system may be divided into four broad sections: (1) Transmission and signaling. (2) Register. (3) Supervisory. (4) Line information and control stores and translation.

Operating speeds of section 2 and 3 are isochronous to section 1, and section 4 is asynchronous to the others. Full details are beyond the scope of this paper, but an endeavor has been made to cover the salient points of the system.

TRANSMISSION

A 4-wire system was adopted because with it a larger number of channels could be accommodated on unidirectional highways, and smaller transmission losses would obtain in traversing the exchange. This latter point is of more importance when a complete network of electronic exchanges is considered, and account must be taken of the stability requirements when one termination is broken. Yet, while "singing" caused by random variations is avoided, sufficient gain must be included in the normal connection to provide transmission quality compatible with the electromechanical network.

Pulse amplitude modulation (PAM) was chosen as the method of transmission on the common time-shared speech paths. It was considered compatible with other control techniques; also it had been shown, on previous experimental models, to give a satisfactory transmission performance. Based on experience from these models, crosstalk and other transmission requirements could be met only with the devices available at that time, particularly diodes, by limiting the packing of highways to 100 channels. Each channel is sampled at a 10-kc/sec (kc per second) rate, which results in a 1-μsec (microsecond) period for each sample. The sampling rate was chosen to suit the transmission of audio-frequencies up to 4.2 kc/sec, a cutoff frequency considered adequate for local speech transmission and also for long-distance transmission with an outband signaling channel.

The lines are divided into groups, each line being capable of connection, by gates, to its respective group IN and OUT 100-channel multiplex highways. To avoid unacceptable side tone, antiphase channel pulses, spaced 50 μsec apart, are used for IN and OUT speech respectively. The group highways are interconnected IN to OUT and vice versa, both within the same group and between groups, by inter-group gates which are controlled by their respective frames of 100-channel pulses. This procedure, known as the switched-highways technique, makes possible a maximum of 50 conversational paths within the same group, or 100 conversational paths to other groups. The total traffic-carrying capacity of the group highways has been calculated to be

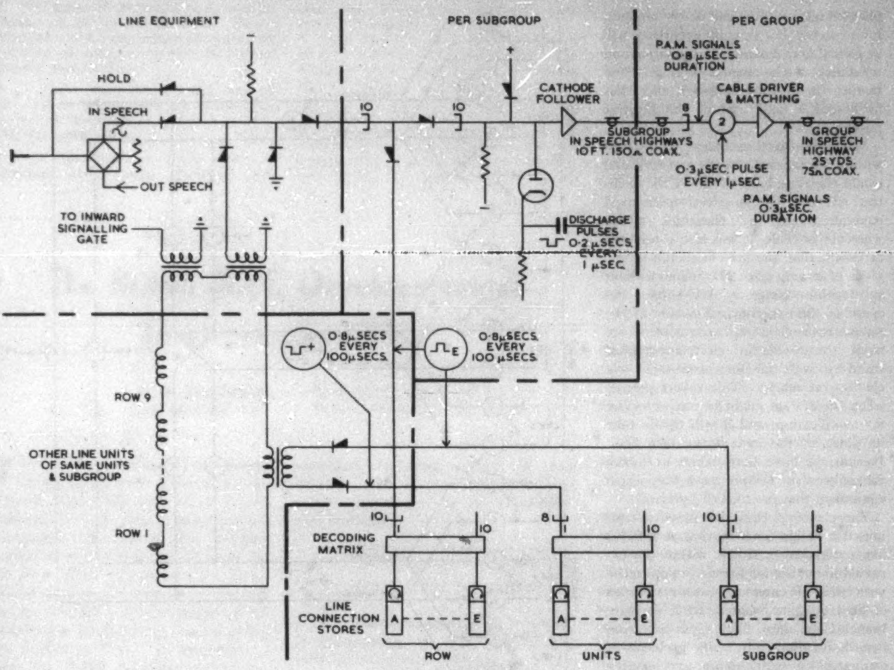

LINE EQUIPMENT
PER SUBGROUP
PER GROUP
HOLD
IN SPEECH
OUT SPEECH
TO INWARD SIGNALLING GATE
10
10
8
CATHODE FOLLOWER
SUBGROUP IN SPEECH HIGHWAYS 10 FT. 150Ω COAX.
P.A.M. SIGNALS 0·8 μSECS. DURATION
CABLE DRIVER & MATCHING
2
GROUP IN SPEECH HIGHWAY 25 YDS. 75Ω COAX.
0·3 μSEC. PULSE EVERY 1 μSEC.
P.A.M. SIGNALS 0·3 μSEC. DURATION
DISCHARGE PULSES 0·2 μSECS. EVERY 1 μSEC.
0·8 μSECS EVERY 100 μSECS.
0·8 μSECS. EVERY 100 μSECS.
E
ROW 9
OTHER LINE UNITS OF SAME UNITS & SUBGROUP
ROW 1
DECODING MATRIX
10
1
10
8
1
10
10
1
8
LINE CONNECTION STORES
A
E
A
E
A
E
ROW
UNITS
SUBGROUP

Fig. 2. Line gates and modulator

approximately 60 erlangs so that the number of lines connected to each group will depend upon the calling rate of the lines. Highgate Wood accommodates a large proportion of low calling-rate subscribers so that each group has a capacity of 800 lines made up of 700 subscribers or outgoing junction line equipments and 100 incoming junctions. Fig. 1 shows the basic trunking with typically three lines in each of three groups. The channel pulses allocated to four conversations are shown:

Connection					
Line No.	Group No.		Line No.	Group No.	Channel Pulses Used
1	1	to	2	1	*P1, P51*
3	1	to	9	3	*P7, P57*
4	2	to	8	3	*P1, P51*
7	3	to	6	2	*P1, P51*

It is seen that the three connections using *P1* and *P51* do not interfere with each other because the gates interconnecting the unwanted pairs of highways are not open at these channel times. The channel pulses to be used on a connection are allocated by a common channel selector from the coincident free channels in the originating and terminating groups

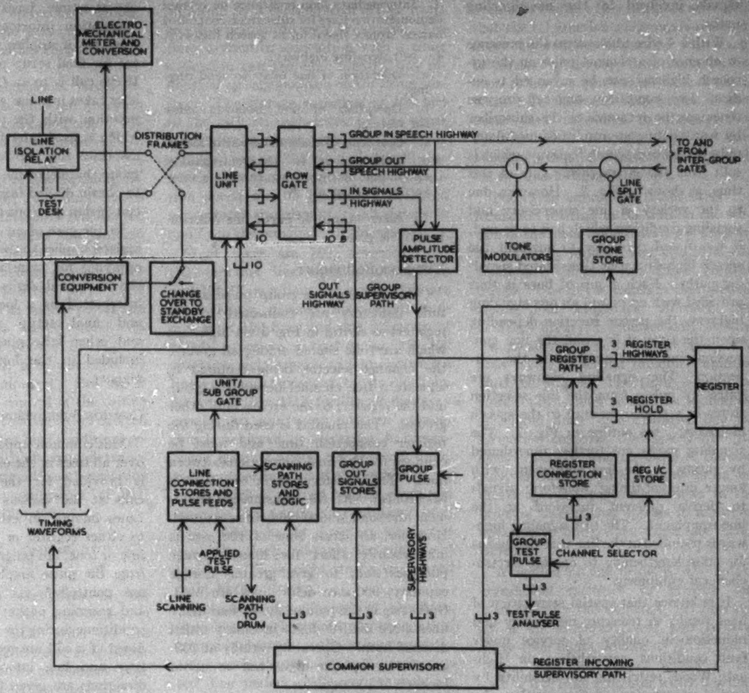
From later description it will be obvious that the progress of a call may be divided into two sections:

1. Initial setup between the calling line and a register. This involves only the coincidence of a free IN speech and IN signaling highways channel pulse of the originating group with a free register channel pulse.
2. Final setup between calling and called lines. This involves the coincidence of free antiphase channel pulses in the respective IN speech highways.

The channel selector is enabled to distinguish between these two conditions, and in practice is fed, in the second case, with the busy channels' indication of the IN speech highway of the calling group and the busy channels' indication of the OUT speech highway of the called group. A sequential scan for coincident absence of these signals results in a free channel indication *Pf*, and the writing of *Pf*+50 and *Pf* into the line connection stores of the calling and called groups respectively. Sequential scanning helps to reduce blocking by the allocation of channels in the same preferred order in each group. The disadvantages of consistently selecting a faulty channel do not apply because all channels in a group will be affected by a fault in the common multiplex path.

The connection of 800 lines to a common highway could be effected by connecting 800 gates to a common point but this would be impractical. To reduce effects of stray capacitance and uneven loading on pulse feeds, the groups are subdivided into subgroups of 100 lines and these subgroups are further divided into rows of 10-line units. A line equipment position is therefore defined by unit, row, subgroup, and group, but since groups have independent sections of equipment (i.e., highways), channel pulse identity is not concerned with the group identity. The channel pulse allocated for the IN speech connection is stored as the memory for the connection in the line connection stores. The stores consist of banks of

Fig. 3. Signaling paths


ELECTRO-MECHANICAL METER AND CONVERSION
LINE
LINE ISOLATION RELAY
TEST DESK
DISTRIBUTION FRAMES
LINE UNIT
ROW GATE
10
10
10 B
GROUP IN SPEECH HIGHWAY
GROUP OUT SPEECH HIGHWAY
IN SIGNALS HIGHWAY
TO AND FROM INTER-GROUP GATES
I
LINE SPLIT GATE
PULSE AMPLITUDE DETECTOR
TONE MODULATORS
GROUP TONE STORE
CONVERSION EQUIPMENT
CHANGE OVER TO STANDBY EXCHANGE
OUT SIGNALS HIGHWAY
GROUP SUPERVISORY PATH
GROUP REGISTER PATH
3 REGISTER HIGHWAYS
3 REGISTER HOLD
REGISTER
UNIT/SUB GROUP GATE
LINE CONNECTION STORES AND PULSE FEEDS
SCANNING PATH STORES AND LOGIC
GROUP OUT SIGNALS STORES
GROUP PULSE
REGISTER CONNECTION STORE
REG I/C STORE
3
CHANNEL SELECTOR
GROUP TEST PULSE
3
TEST PULSE ANALYSER
SUPERVISORY HIGHWAYS
TIMING WAVEFORMS
APPLIED TEST PULSE
LINE SCANNING
SCANNING PATH TO DRUM
3
3
3
3
COMMON SUPERVISORY
REGISTER INCOMING SUPERVISORY PATH

100-μsec recirculating magnetostrictive delay lines which circulate the line's code in the particular channel time. Center-tap outputs from the delay lines are also provided to give the antiphase pulse required for the OUT speech connection.

The identity of each line equipment within a group is defined from three banks of five delay lines using a 2-out-of-5 code for the units, row, and subgroup with a decoding matrix to give an output in decimal form in each case. These outputs are connected to the IN speech line gates as shown in Fig. 2, and a similar arrangement is made from the antiphase output to the OUT-speech line gates. Thus, with the knowledge of a line equipment number, the corresponding 3×2/5 code can be inserted in the line connection store at the particular channel pulse time P of the 100-channel frame. PAM signals will be transmitted into the group highways from the line at that time, and to the line in antiphase, i.e., at $P+50$. The other line of the connection having $P+50$ for the IN-speech, and hence P for the OUT-speech gates, allocated in its own code at the same channel time, will provide a conversational path.

It will be noticed from Fig. 2 that pulses of 0.8 μsec are used in the modulator. They have an unmodulated height of 8.5 volts with a maximum modulation of ±3 volts superimposed by the IN speech. Losses through the system give a resultant output height of between 4.5 and 7.5 volts which is adequate for demodulation. A discharge gate is provided prior to the subgroup highway to overcome gating capacitance. Prior to the group highway the amplitude-modulated pulses are reduced in duration to 0.3 μsec because of unacceptable crosstalk if less than a 0.7-μsec channel spacing is used in the 25 yards of cable forming the highway. This cable length is determined by maximum apparatus disposition and is made constant to effect constant delays in the system. Pulses are connected on the highway only if the line-hold condition is present. This facility is incorporated to facilitate supervision and also to prevent singing if lines present an open-circuit condition.

The OUT-speech circuit consists of a pulse-lengthening circuit from the group highway to restore the group pulses to 0.65 μsec, followed by cathode followers to drive the subgroup highways. The subgroup highways feed line gate arrays driven from the antiphase decoding matrix but otherwise similar to those found on the IN-speech highways. These are followed by a capacitor pulse lengthener and low-pass filter which form the demodulator, and a line amplifier feeding the line hybrid.

The intergroup gates are driven by separate and individual single-output magnetostrictive delay lines since no advantage could be gained by coding the intergroup connection stores for a small number of groups.

SIGNALING

In any common control system the signaling involved between the various blocks of equipment is generally complex, and this applies equally to this system. It is therefore possible to deal with only the general outline of some of the signals involved in the line-signaling paths.

With a 4-wire tdm system the presence or absence of a channel pulse on the IN-speech highway can be arranged to indicate loop conditions, and call progress tones can be returned to the subscriber by transmitting an amplitude-modulated pulse on the OUT-speech highway, which is split from the interhighway gate at this time, as shown in Fig. 3. However, due to the variety of line supervisory and signaling conditions which had to be met, it was found necessary to augment the speech highways with time-shared signaling paths. Each group of lines is thus provided with an IN and an OUT signaling highway, the precise function depending upon the type of line and the line conversion equipment involved in the connection. The signaling highways are gated by pulses from the line selection stores in a similar manner to the speech highways; see outline in Fig. 2. The signaling paths are further time-shared by gating, in the line equipment, with longer interval-timing waveform signals to derive different functions at the line equipment. The OUT signaling highway is provided at the line with the means for pulse lengthening to operate electrochemical equipment.

It is known that spatial segregation of lines results in trunking inefficiency and reduction in quality of service under fault conditions; nevertheless, on Highgate Wood, restricting full flexibility by partial segregation of lines within the groups has proved to be economical. Thus, while an ordinary subscriber may take up any equipment position, party lines, incoming and outgoing (IC and OG) junctions, and PBX lines are restricted to particular subgroups. By this means, since ordinary subscribers, PBX, and IC junctions do not require inward signaling paths, only two subgroups are fitted with IN signaling paths.

Fig. 3 shows the group-signaling highways and associated equipment. The pulse amplitude detector connected to the IN speech and signaling highways indicates to the exchange controls the amplitude range of each pulse on the highway. Some of the facilities provided are as follows:

1. Discrimination between shared-service calling line of X or Y party by the appropriate signals on the IN speech and IN signaling highways when the subscriber grounds the appropriate leg of the line to call into a shared-line circuit.
2. Called subscriber answer and manual-hold indications from outgoing junctions via the IN signaling highway. This is used to detect full-loop or zero conditions only.
3. Intermediate loop resistance on certain demonstration lines for subscriber-controlled barred trunks signal on IN speech highway.
4. OUT signaling highway:
   a. Operation of line relay to send ringing.
   b. Operation of line isolation relay for testing.
   c. Operation of electromechanical meter.
   d. Pulsing out to electromechanical exchanges by operating relays in conversion equipment.

In cases 4a and 4c party line discrimination is given.

REGISTER EQUIPMENT

Outputs from the group pulse amplitude detectors are connected to the registers as shown in Fig. 3 via highways which are time shared among all groups, the channel selector being required to allocate a free channel between a group and the register, to the exclusion of other groups. This channel is used during the register connection time and must be changed for the final connection between lines. The register system operates in isochronism with the transmission equipment and some similar delay lines are used. However, an access time of 100 μsec is unnecessarily short for many storage purposes and, to give greater storage capacity, 900-μsec delay lines are used. Operating in the parallel mode, four delay lines store dialled digits in binary-coded decimal form. Operating serially at 100-μsec intervals, nine digits can be stored for each of 100 registers.

Traffic at Highgate Wood requires only 22 registers. It is possible to time-share register highways to perform separate functions and reduce the number of registers with saving in storage apparatus. The system however, uses alternate channel pulses of the transmission system, thereby halving the clock frequency of the register equipment, i.e., reduced to 500 kc/sec.

The register acts on originating calls from subscribers and junctions. It receives information, via staticisors, regarding the class of service of the calling line, and subsequently of the called line, from the line information store, so that indications can be given to the common supervisory over separate time-shared signaling paths to actuate the appropriate tone stores and line-operating conditions.

The action required of the register is determined by the class of service, e.g., IC junctions will send only numerical digits, or disabled subscribers who cannot dial will be routed automatically to the parent manual board. When sufficient code digits have been received, the register makes application to the translator on a one-at-a-time basis via staticisors for translation information. When this is received, it applies to the line comparator for final setup procedure of the call. If the call is to an OG junction, the register takes into use a register link, in conjunction with the common supervisory, which is used later for pulse changing to associate the register channel with the group channel of the OG junction in order to obtain dialled impulses at the junction conversion equipment.

In the case of an IC junction call, the register applies to the comparator without applying for translation and only with the numerical digits. This condition is interpreted as a local terminating call, and final setup proceeds. Time-out and other false operation protection is included in the logical routine of the register.

COMMON SUPERVISORY

This common apparatus is time-shared over all lines in the exchange, and storage is provided for the supervision of all calls at the various stages of progress. Lines having an established connection to either register or other terminal, i.e., line or tone, are progressed via the signals from the pulse amplitude detectors and are controlled via the group-signaling and scanning paths; see Fig. 3. Other conditions during the setting up and clear-down of a call are received from the register and line information stores, and directions are given to these equipments for correct progressing of the call over separate time-shared signaling paths. The supervisory is separated from the register to avoid wastage of register equipment which is required only during the initial setup of the call.

The initial system design requirement was to cater for nine groups of 100 channels each, so the storage devices used in the supervisory are 900-μsec recirculating magnetostrictive delay lines. Signals between the supervisory and the group equipment are interlaced; to each group is allocated a train of pulses which occupy every ninth channel pulse position. In this way each group 100-channel frame is scanned completely in 900 μsec by a vernier system. However, the information derived from a scan and relating to a channel in a group must, in general, be used at the antiphase channel time in another group, e.g., for information which has to be exchanged between lines. A time-shared intergroup signaling path is provided for each group into which the supervisory writes information and receives it back in antiphase gated by the appropriate group pulse under control of the intergroup connection stores. This exchange path can pass information, such as ANSWER, in either direction once every 2 millisec.

The supervisory equipment stores information relating to allocated channels both on the calling and called side of the connection, as well as to relevant class of service.

Persistance checks of line-signaling information are made; auxiliary stores carry out timing operations to give limited periods of tone application and junction guard. Local call-charging on a pulse-on-answer plus random-point-of-entry into a common repetitive pulse cycle basis is effected. The meter pulses are passed to the line meter stores via the scanning path shown in Fig. 3. Clearance conditions are controlled by the supervisory equipment; clearing is normally under the control of the calling line. Manual-hold conditions are catered to. Called subscriber held conditions result in a time out with ultimate clearance of the calling line onto a "park" condition after intermediate application of NU (number unobtainable) tone.

Line Information, Control Stores, and Translation

In conventional exchanges of the step-by-step type, the association of lines from their cable order to the directory number order, and from directory number order to equipment position number order is accomplished by wire-jumpering on what are commonly known as main and intermediate distribution frames (MDF and IDF) respectively. The additional facilities offered by an electronic dynamic store, including the replacement of the IDF, led to the use of a magnetic drum to perform the functions covered in this section.

Initial design caters for complete flexibility in the association of directory and equipment numbers but at Highgate Wood other factors, as previously noted, have led to the partial segregation of lines within groups. It was necessary in this installation to retain an IDF because of the association with the electromechanical stand-by exchange and because it is required to operate the subscribers' electromechanical meters from either exchange, as well as to perform the metering function electronically. The main distribution frame has been retained for its protection facilities and also for connecting some of the conversion equipment, as shown in Fig. 3.

The magnetic drum has a number of tracks, each of which is divided into 100 words, each of 28 binary bits, which are indicated by a clock system driven from, and in synchronism with, the drum. The clock system in association with a track indication can therefore identify any individual storage position on the drum surface. The storage is packed at 100 bits to the inch on each track by the phase modulation method; the drum speed is arranged so that the bit period is 10 μsec which gives the associated logic a repetition speed of 100 kc/sec. This speed necessitates staticization of all signals to and from the remainder of the system.

Since there are 700 subscribers' or OG junction line equipment positions and 100 IC junctions per group, they are conveniently divided into seven subgroups of 100 lines and one subgroup of 100 IC junctions so that each of the 100 words on eight tracks is identified with a line equipment position by its clock indication. The IDF facility of directory number to equipment position association is achieved by recording, in binary-coded decimal digits, the required directory number into 16 of the 28 bits of the word which is defined by the clock and track as the associated equipment position. Thus, as the directory number is read, the clock indicates the equipment position. These eight tracks are termed main library tracks.

Detection of a line-calling condition, from lines which are not already marked busy, is achieved by allocating one of the 100 speech channels in each group as a test line period; different channels are allocated in different groups. Within each group the channel is cyclically allocated by the line equipment clock information from the drum to each line equipment, which provides sequential line scanning. This information persists for 1-word time, i.e., 280 μsec, and is in the form 3×2/5, which is identical to that used in the line connections stores through which it is applied to the line; see Fig. 3.

A full-amplitude pulse will appear on the group IN-speech highway if the tested line is calling. This pulse is passed via the pulse amplitude detector and group test pulse gate to a common test pulse analyzer. The analyzer detects a "connection to register required" signal and identifies the calling group to inhibit its scanning until the relevant information is obtained from the line store on the drum. The channel selector then effects a connection to the register.

Simple cyclic scanning of eight subgroups would result in a scanning repetition time of 224 millisec which is too long for IC junctions, apart from the additions of scanning inhibition times of approximately 8 millisec, while an IC or OG connection is made. It is therefore arranged that the IC junction subgroup is scanned every 28 millisec in parallel with each subscribers' subgroup, then finally on its own, to make a convenient cycle of eight scans. To facilitate staticization it has been found expedient to allocate two channel pulses in each group for line scanning; thus 98 channels are employed for speech at Highgate Wood.

The drum equipment stores other information relating to the lines in the section of the main library word not used by the directory number, and also in words positioned similarly on three parallel tracks. One of the tracks is time-shared among four subgroups. Another, the auxiliary library, is required only on subgroups serving shared service and PBX lines where a second directory number must be stored. The third is used to store totaled meter records for individual or shared lines. Miscellaneous line information concerns the type of line, its class of service, and the state of the line such as busy or park, as well as an indication of the sequence number in the order of testing that applies to multiple-appearance lines when these are being called.

A section of the drum is devoted to translation facilities to provide one translator for the exchange. It operates on a one-at-a-time basis and receives three code digits, known as A, B, and C digits, and the class of service of the calling line from a register which has been allocated the translation signaling paths at a particular time. The translator scans over the one track (of ten tracks) which corresponds to the A digit, to read off the translation information corresponding to the B and C digits. This drum word on the track is indicated by clock signals which correspond to the one (out of 100) indication of the B and C digit combination. The translation information read off the 28-bit word is passed back to the register and may include four routing digits to control the route of the call in the external network, in addition to a local address which contains the information for OG call routing within the exchange. The type of connection, local, OG direct or tandem, or national is also indicated.

When a condition for the final setup of a connection exists, the register is allocated to the drum equipment, which is now used to perform a line comparator function on a one-at-a-time basis. This condition exists when a translation has been received for OG calls or when the complete directory number is available for locally terminating calls. The information as to the required line, directory number or junction, or service route number, is passed to the comparator where it is compared with the output of all library tracks in parallel.

Hence in one revolution of the drum (28 millisec) the line is found, and from its track and clock position the group and equipment position is determined. This information together with relevant line information read off the drum is staticized and the line marked busy. When the system is not making an initial setup to a register, the called line's group and equipment position in 3×2/5 form is passed to the channel selector together with an indication of a terminating call. Thus line-scanning and line comparison proceed independently; their only interference occurs during the short time when access to the channel selector is required.

On receipt of terminating call information, the channel selector obtains information as to the calling line's equipment position from association of the register channel in use and the register connection store; it then performs the selection of a suitable channel for final connection setup, as previously noted. The register is released and the call controlled by the common supervisory.

In the case of line comparison to a multiple-appearance PBX line or junction, the sequence number of all free lines in the route is noted and the lowest number chosen for connection. This is achieved by successive deletion of the staticized information as a free lower sequence number is encountered during one drum revolution. A conditional busy applied to all lines in the route during the selection is removed during a subsequent revolution.

The transfer facility, which provides for rerouting a terminating call to a third party, is catered to by drum tracks into which the directory number of the transferrer and transferee are written. With the indication of a transfer in operation in the library, the directory number of the transferee is determined and the call set up as if this were the required number.

If the called line is not found during a revolution of the drum or if it is found to be busy, arrangements are made to give NU or busy tone from the calling line's group tone stores, and the register is released.

The system offers many other standard facilities and some new ones readily provided by the trunking and by electronic means. The security techniques which are based on duplication vary among the different parts of the system, and valuable experience is anticipated from their operational results.

Operation of Full-Scale Model Exchange and Design of Highgate Wood

The small differences between the two designs are due mainly to the inclusion, in the Highgate Wood design, of security methods to maintain service in the face of a fault in common equipment and the provision of maintenance aids. In addition, the design of the field trial includes only those facilities suitable to a director-type exchange in the British network.

The model has been in operation at the British Post Office Research Station since early 1959; its chief function has been to prove the practicability of the system, and to enable individual sections to be tested, both separately and in association with their related equipment, on an equipment scale representing actual exchange practice.

The commissioning and operation of the model has been subject to difficulties based on problems in bringing together personnel who, while they are specialists in their own sections, are not initially fully conversant with the details of the other sections with which they interwork. The experience gained has resulted in improvements to the Highgate Wood design and stimulated the research effort for future systems. Transmission test results give adequate stability and are generally better than the target design which for the major factors was as follows:

1. Over-all loss in modulation level (speech component) in common equipment 0.5 db (decibels) ±0.5 db.
2. Crosstalk attenuation between any two connections >70 db.
3. Permissible distortion of speech signals
   a. With 4 milliwatts incoming from 2-wire line (represented by ±3-volt modulation) <5%.
   b. With 1 milliwatt incoming from 2-wire line <1%.
4. Permissible noise level at 2-wire point < 0.63-millivolt psophometric electromotive force.

The crosstalk figure is satisfied with the use of nickel-doped germanium point contact diodes and channel sample spacing of 0.2 μsec to cover hole storage times. An individual circuit attenuation of better than 90 db is attained. However, it does appear as though 100 channels in a single multiplex is the practical maximum from economic considerations. This is due mainly to the cost of commercially available devices and to the requirements of economic equipment design in regard to restrictions on stray capacitance variations, and hence delay variations, within a clocked system.

During the time of development, the availability of fast-switching transistors, at competitive prices, has permitted their use in alternative techniques which avoid the use of hot cathode valves. It is therefore known that saving in power requirements, and hence space, are possible in future high-speed tdm systems.

The control techniques using computer-type logic in association with magnetostrictive delay lines and magnetic drums proved satisfactory. The asynchronous conditions, however, involved costly staticization and complexity which should if possible be avoided in future designs. Following the implementation of a unified design specification and avoidance of a multiplicity of power supplies, a practical high-speed tdm system has proved realizable. Full reliability information will be obtained from continued use of the model and operation of the field trial exchange.

It has always been recognized that the extensive use of common apparatus necessarily increases the risk of system breakdown arising out of a component or circuit failure, and from the outset the need was foreseen for organizing the system to reduce the risk of system breakdown to an acceptably low value. Lack of field experience was, however, a limiting factor and the approach to system security in the Highgate Wood exchange has been necessarily experimental. With the exchange design now complete a better understanding of the factors affecting continuity and quality of service is obtained, particularly in the study of system economics. It may therefore be expected that subsequent designs will depend on a modified system organizaion aimed at securing a high standard of performance rather than a minimum of time-shared control apparatus.

Resulting from an appraisal of the Highgate Wood design, still further economies are under consideration, such as a reduction in the flexibility given by IDF facilities but, as in all cases, the effect of ultimate operating costs must be estimated.

In general, therefore, the flexibility of a practical tdm system in providing a wide range of facilities and satisfactory transmission has been proved. The necessity for simplification from the sophisticated design to obtain satisfactory maintainability and also an economic system has been determined. Security must be inherent in the system design and consideration must be given to the desirability of meeting the full range of exchange sizes with one standardized system. With the continued reduction of component costs and economies effected in the light of the foregoing considerations, a competitive electronic exchange system should be possible.

Compatibility with Existing Systems and Future Developments

Not the least of the problems associated with electronic exchanges are those of introducing them into well-developed national systems comprising electromechanical exchanges with peripheral apparatus designed to suit these exchanges. It is inevitable that for many years every new electronic exchange will find itself compelled to communicate over junctions with, and to take over some lines from, existing types of exchanges. Interworking over junctions between different types of apparatus is a question of conversion equipment, usually in the new exchange but possibly in either or both exchanges. The technical problems are not difficult; the expense is unfortunate, but moderate. Subscribers' sets working on direct exchange lines and PBXs connected to public exchanges operate with line currents of 30 to 100 milliamperes which suit the relays in the exchanges. Similar matching of subscribers' set and electronic exchange apparatus would result in 10 to 20 milliamperes of line current, and there is every inducement of cost and power supply to establish a new practice of this order. Existing stations are rung by low-frequency a-c signals of up to 1 watt in power, to ring bells and to drop indicators. The kind of signal which electronic exchanges can most economically emit for these purposes is an audio-frequency of a few milliwatts in power. The alternatives presented are station equipment suited to the ringing capabilities of electronic exchanges or a relay per line to send the existing low-frequency signal, with reduced line-feeding current as an additional problem. At present, both of these solutions are expensive; they are being studied for a long-term solution.

A more immediate problem concerns what to do with already existing plant which suddenly is included within an electronic exchange area. In the previously mentioned experimental exchange at Highgate Wood the problem is temporarily avoided by the condition of the experiment, that it must be possible to operate all the lines from the electronic or a regular exchange. To this end the electronic exchange carries extra, costly equipment which is justified by the experimental nature of the exchange, rather than by the economics.

It is presently believed that apparatus suitable to electronic exchanges will eventually be produced; the intention is to design electronic exchanges to accommodate both existing station apparatus and low-feed-current tone-ringing apparatus.

All service in the United Kingdom is metered at the exchange. In addition, customers may also have meters installed on their own premises if they wish to ascertain charges as they are incurred. The exchange meters are read quarterly in the accounting system. Although electronic exchanges can provide these facilities in their present form, a record is planned for the future which can be read automatically to facilitate machine accounting, and which will be more suited to electronic exchanges than individual meters in the exchanges. Tape recording, magnetic drums, and ferrite matrices are possibilities for storage of the meter record.

In the existing service certain operational and maintenance practices must be continued for an electronic system at least until experience shows alternatives more suited to electronic working. For example, exchange lines can be tested for insulation resistance up to 2 megohms from a remote test desk. This requires d-c isolation of the line from the exchange equipment, a condition which naturally occurs with relays in exchanges. In electronic exchanges it requires the provision of relay contacts having no other function, constituting an economic handicap. Providing service observation and trunk offering (i.e., "butting in" by an operator on an established call) can also be difficult and costly in relation to the infrequency of their use on any one line. For the Highgate Wood experimental exchange all existing facilities are being provided; it is hoped that alternatives may be found where the provision of a facility is an economic disadvantage to the electronic operation.

Space-Division Systems

In space-division systems each connection path occupies a unique position in space. Interconnection is accomplished by switches termed "cross points" from their determination within an array by the intersection of co-ordinate markings. In this sense electronic space systems are similar to crossbar systems but the speed of operation of electronic cross points, and their control, permit a high degree of common control. A well-known technique for a number of systems which has been investigated is the control of calls on a one-at-a-time basis using end markings of a multistage switching system with route selection through the stages.

In the investigation of electronic space systems, the cross points, because of their profusion, were found to be the most significant factor in the determination of the economics of the system. There may be typically 25 cross points per line and they may account for up to half of the total cost in a competitive system. They also determine, to some extent, the signaling method used between lines, since in most cases for economic reasons an earth-return (unbalanced, single-wire) system has been considered, and transformers are generally involved as the linking medium between switching stages. Time-shared auxiliary signaling paths have been devised but they generally overcomplicate the system.

Cross points which have been investigated include gas tube diodes, gas tube triodes of the single- and double-anode type of semiconductor diodes, transistors, thyristors, and p-n-p-n semiconductor diodes. While it was possible to construct model systems using each device, they all suffered shortcomings, either technical or economic.

Economic possibilities may still exist for a few of the devices, and the situation is under review. In all cases, attainment of the most efficient trunking and selection of path control methods is essential. The advantage of electronic cross points over crossbar switches lies in their flexibility of switch size in any stage and in their compatibility with electronic control methods. In addition, when sealed, they are generally unaffected by ambient atmospheric conditions and no mechanical wear is involved in their operation; their life expectancy should therefore satisfy exchange requirements of longevity.

Following the development of the dry-reed relays by Bell Telephone Laboratories and their acceptance as electronic components, effort has been applied toward developing their use for a cross point in electronic space systems. They have the advantage of electronic components in being sealed and, because they may consist of four make reeds, when assembled in a cross point, they can provide a balanced speech path (2-wire), a through-signaling connection (1-wire), and a holding contact. They therefore form the basis of a system which can be completely compatible with existing electromechanical systems without incurring complex conversion equipment. Life

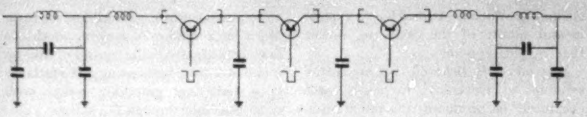


Fig. 4. Two-wire resonant transfer

tests on reeds give every indication of satisfactory performance.

The operating speeds of reed cross points still permit one-at-a-time operation of connection paths, and they may be used in a system using trunking methods considered for other types of cross points. The main consideration is again economic, but with fully automated production of the reed inserts it is hoped that reeds having noncritical operating margins can be made at a cost which will permit a competitive system for all but the very large exchanges.

Mk II TDM Systems

Two-Wire Resonant Transfer Systems

The basic transmission conditions of a connection through an exchange using resonant transfer 2-wire tdm transmission through a number of switching stages is shown in Fig. 4. Inevitably, included between the two terminal ends, are short lengths of transmission medium and multipled gate circuits. These must not appreciably degrade transmission or produce crosstalk to other circuits. There is no great difficulty in including the intermediate transmission paths in the resonant transfer with moderate attenuation as its only effect; this and presently available components limit the multiplex to about 24 channels at the 10 kc/sec (kc per sec) prf (pulse repetition frequency) which is contemplated in all tdm systems. Crosstalk minimization requires careful mechanical construction and some limit to the size of any one switching stage. The maximum practical size has not been determined accurately; it is estimated as a 25 by 25 switch.

The ~~mean loss through an~~ exchange with resonant transfer transmission may be from 2 to 4 db depending on the funds to be spent on components. The random variation of loss about the mean is small.

Four-Wire Systems

The electrical characteristics of transistors favor "constant-current" operation of 4-wire multiplex transmission. This means a datum current modulated by a speech signal and transmitted through transistors switched by constant voltage pulses. Fig. 5 shows in outline what is now contemplated. The modulator consists of a transformer-inductor *L1*, capacitor *C*, and diode *D*, fed by a datum bias current and signal current and a gate, *G1*. The datum current as shown is the loop current; the modulating current, the audio current received from an audio line via a hybrid coil; and the gate, a transistor. When the gate is operated by a voltage pulse (applied to the base of the transistor), the instantaneous current in the primary of *L1* is transferred to the secondary and thence via the transistor gate *G1* and other gates operated by pulses of increasing voltage to a demodulator.

The demodulator shown comprises an inductor *L2*, which is charged by the current pulse received through gate *G2*, and subsequently discharged into a transistor. This produces a current pulse into the filter *WF* to complete the demodulation, the audio output being supplied via the hybrid to the audio circuit. Presently available components set a limit of about 100 channels in a single multiplex (at 10 kc/sec prf). Because of the individual amplifier the mean loss can be zero; the variation of mean loss may have a standard deviation of up to 1 db if normal tolerance components are used.

Systems with a smaller number of channels become easier to design and less exacting in the components required. Four-wire 24-channel multiplexes are little more expensive than the 2-wire resonant transfer type and have a lower mean loss. Moreover, the unidirectional transmission paths are less subject to crosstalk; hence larger switches are possible with 4-wire than with 2-wire multiplexes, which makes them attractive for larger exchanges.

Network Planning

For many years, the possibilities of far-reaching changes in exchange area planning brought about by electronic exchanges have been contemplated. No great progress was possible until the shape of electronic exchanges could be seen in some detail, and this has only recently been the case. The problem resolves itself into two aspects: improved transmission, and economy in line construction.

Improvement in Transmission

In their transmission characteristics, 2-wire switched electronic exchanges of the space- or time-division type are not basically different from metal contact 2-wire switching exchanges. Their practical performance is likely to be a little worse because of a slightly higher loss through the switches and also because of impedance irregularities not present in metal contact exchanges in passing through the switches. Four-wire transmission requiring exchange lines to be connected to the exchange

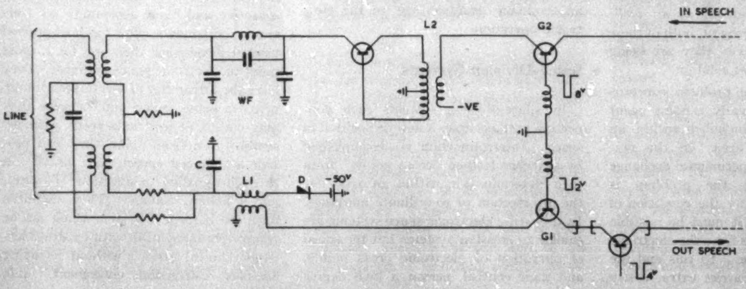


Fig. 5. Current modulation system

equipment via 2- to 4-wire terminations can also degrade transmission if low balance return losses occur at any frequency within the speech band. The degradation would consist of attenuation and attenuation distortion (decreasing with increasing attenuation) due to the circulating path within the exchange when two such lines are connected together. The opportunity for improving the existing analog transmission and switching networks with both existing practices and electronic exchanges depends on reducing the impedance mismatches and internal reflections which occur at the switching points. The necessary improvement would be realized with amplifiers at the terminal ends of all junction circuits, a situation which has so far been economically impossible. Four-wire tdm exchanges would provide this situation without additional expense, a fact which reacts favorably on the economics of this kind of exchange.

Economy in Line Construction

If it is economically possible to include gain in a local connection (and with 4-wire tdm exchanges this is not difficult) then cheaper local line construction will result. Means of varying the gain or loss of connections through local exchanges are also being investigated whereby gain can be inserted in calls which would otherwise involve too much loss, and attenuation in calls which would otherwise be too loud or unstable.

Pulse code modulated transmission between concentrators, as suggested by the Bell Telephone Laboratories, or between terminal exchanges with analog local transmission is being investigated.

Future Plans and Conclusions

A 100-channel 4-wire tdm exchange will be factory-manufactured in 1961 and is to be installed at Highgate Wood in the London area in parallel with a new normal program exchange. When commissioning tests have shown the electronic exchange to be suitable for public service it will be cut over and cut back only if some major difficulty arises. An Mk II system in service operation by 1965 is planned. The trial of other than 4-wire tdm types of exchange is dependent on the results of the studies now taking place.

The main conclusion at present is that a number of techniques exist by which practical and economic electronic exchanges may be realized. The outstanding problems are those of reliability and security. Reliability is measured by the rate at which faults appear and have to be cleared. Over the whole range of electronic components and taking into account their relative numbers, a fault rate of about one fault per 2,000 components per annum is currently being quoted. With an average of 50 components per line, this means one fault per 40 lines per annum, which is comparable with electromechanical system fault rates but not good enough for ready acceptance in electronic systems.

Security is measured by the effects of faults on the service to customers, and particularly the denial of service to some or all of the lines on an exchange. The problems tend to be more severe in electronic than in electromechanical systems because of the greater use of common control. Security standards are being devised, e.g., for the maximum time that lines can be out of service because of exchange faults throughout the life of an exchange. Future systems must satisfy these standards.

Two-wire space and low-channel capacity tdm systems are also under development and study for economic assessment relative to the maximum channel capacity tdm systems which so far appear to be the most economical and practical for medium and large exchanges. Small exchanges are being treated as a separate problem and it is here that some of the simpler systems may prove economical.

Effects of Cable Irregularities on the Operation of Telephone Repeaters

S. I. McCARON
MEMBER AIEE

IDEAL CONDITIONS for voice frequency telephone repeater operation consist essentially of having line facilities with infinite "return loss." In practice, these conditions cannot be achieved, nor need they be. All that is actually needed is an adequate degree of balance at all frequencies of interest between two circuit elements, such as a line and its balancing network, to permit deriving the required amount of repeater gain and still to have a 3 to 6 db (decibel) margin of gain before oscillation or "singing" occurs.

Return loss is a measure of the degree of impedance match or balance between two circuits or circuit elements, such as impedance Z_1 and impedance Z_2, and is expressed by

$$\text{return loss} = 20 \log \left(\frac{Z_1+Z_2}{Z_1-Z_2}\right) \qquad (1)$$

where $Z_1 > Z_2$.

From equation 1 it is readily apparent that impedances play an important part in the return loss picture.

The impedance of a cable pair is determined by its four primary constants, namely, resistance R, conductance G, inductance L, and capacitance C, and can be expressed by

$$Z_0 = \sqrt{\frac{R+jwL}{G+jwC}} \qquad (2)$$

where $w = 2\pi f$.

The conductance G in cable circuits will be very small and, since the wires are very close together, the inductance will also be small compared to the other parameters. As a result, both of these parameters may be neglected. Thus, equation 2 for a cable without added inductance reduces to

$$Z_0 \approx \sqrt{\frac{R}{jwC}} \qquad (3)$$

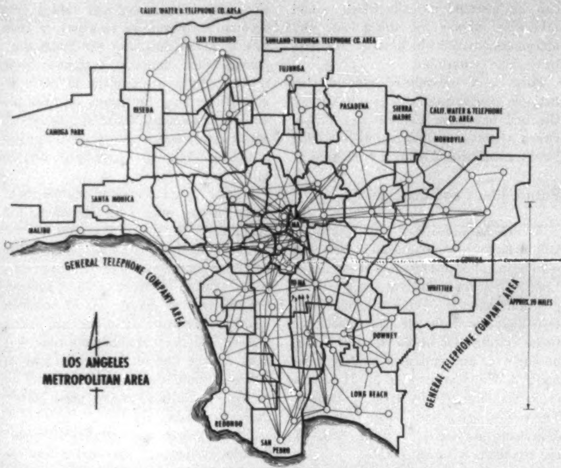
If a cable pair is loaded by inserting inductors in series with the line at uniform intervals, wL becomes very large. In fact, by comparison, R may be disregarded or treated as a constant K in the original equation, which now reduces to

---

Paper 61-118, recommended by the AIEE Wire Communications Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted September 30, 1960; made available for printing January 10, 1961.

S. I. McCaron is with the General Telephone Company of California, Santa Monica, Calif.

The author wishes to acknowledge the work of J. E. Koenig who was responsible for the development of the method used in obtaining the repeater response curves.


CALIF. WATER & TELEPHONE CO. AREA
SAN FERNANDO
SUNLAND-TUJUNGA TELEPHONE CO. AREA
TUJUNGA
RESEDA
CANOGA PARK
PASADENA
SIERRA MADRE
CALIF. WATER & TELEPHONE CO. AREA
MONROVIA
SANTA MONICA
MALIBU
LA
GENERAL TELEPHONE COMPANY AREA
COVINA
WHITTIER
APPROX. 20 MILES
GENERAL TELEPHONE COMPANY AREA
DOWNEY
LONG BEACH
REDONDO
SAN PEDRO
LOS ANGELES METROPOLITAN AREA

The Pacific Telephone & Telegraph Company serves the Central Area

Fig. 1. Map of trunk routes in Los Angeles extended area

$$Z_0 \approx K\sqrt{\frac{L}{C}} \tag{4}$$

From equation 4, it may be seen that the impedance of a loaded cable pair is primarily dependent on the inductance of the loading coils (which are controlled to ±1/2% in manufacture) and the cable capacity per loading section. Since the inductance of load coils is very constant, they will not create impedance variations by themselves. As much cannot be said for capacity variations in paper-insulated lead-covered cable which may vary as much as 14% from minimum to maximum. If tape armor is also placed on the cable, additional stresses during manufacture will be imposed. Also, stresses during the installation process can add substantially to the variation of capacity between pairs.

A major source of impedance irregularities is the more or less accidental variations in cable capacity per loading section resulting from any one or all of the following:

1. Splicing together of cables having different characteristics.
2. Deviations of the coil locations from the theoretically ideal, for various practical reasons.
3. Differences in the interwinding capacity of the load coils themselves, as supplied by different manufacturers.

These capacity variations in the final analysis are the source of most of the impedance variations that determine the structural return loss of the loaded cable pair.

The impedance variations which determine the structural return loss cause a series of internal reflections of the incident energy at the impedance discontinuities. These reflections could be described in terms of individual reflection coefficients and then combined into a single value and referred to either end of a cable. In practice, reflection coefficients are awkward to measure or compute; therefore, in their place, structural return losses are usually preferred. Structural return losses can be readily measured and are simple to relate to reflection coefficients, as shown in the following:

$$\text{structural return loss} = 20 \log \frac{1}{\text{reflection coefficient}} \tag{5}$$

The measured value of the structural return loss over the voice frequency range is ideal as a tool in determining the electrical uniformity of a loaded cable; it is a measure of the exactness with which the network impedance of a repeater and the sending-end impedance of the line agree. The return loss should be as high as possible: values above 30 db are desirable, values above 20 db are usable, but values below this limit may imply that the cable is questionable for some services.

Typical Metropolitan Telephone Network

In a metropolitan telephone network, such as the one in Los Angeles, there are many central offices which need to be interconnected. Fig. 1 shows a map of the present trunk routes for the Los Angeles extended area and the five connecting companies which serve more than 3 million telephones. The General Telephone Company of California serves the western area, part of the southern area, and the eastern area. The Sunland-Tujunga Telephone Company and the California Water and Telephone Company serve part of the northern area while the remaining area is served by the Pacific Telephone and Telegraph Company.

Over the years, the central offices have been interconnected by new circuits as the population increased and the suburban areas were developed. Some idea of the size and scope of this development may be indicated by using, for an illustrative example, the San Fernando Valley which covers the northwest area shown in Fig. 1. During the postwar period this area has shown a population increase from a few hundred thousand to almost one million people. This has necessitated a tremendous increase in new central offices

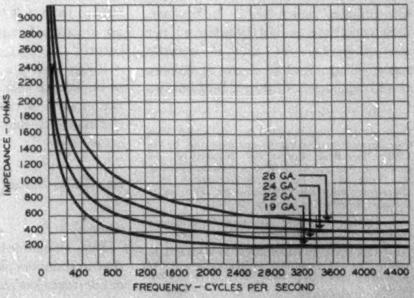


Fig. 2. Typical impedance characteristics of non-loaded cable

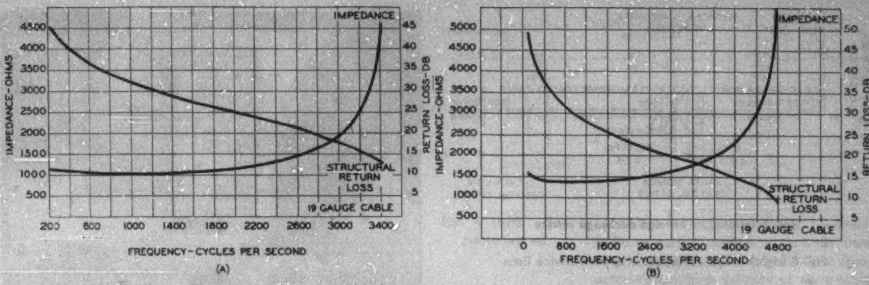


Fig. 3. Impedance and return loss characteristics. A—For H88 loaded cable. B—For B88 loaded cable

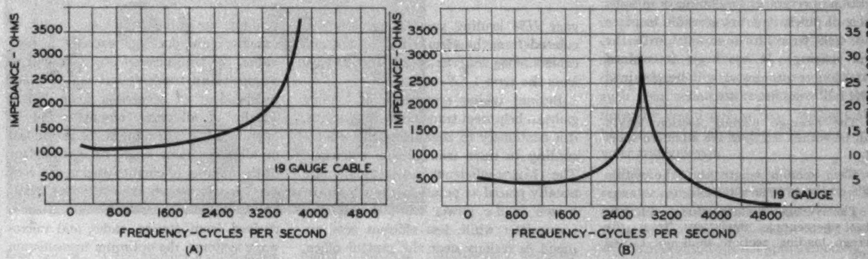


Fig. 4. Impedance curves. A—For D88 loaded cable. B—For M88 loaded cable

and a complex trunking system which has resulted in a variety of types, sizes, and gages of cable with an equally great variety of impedances.

In loaded cables, the spacing between load coils and the inductance of load coils are factors in determining cable impedance characteristics. Cable loading types are designated by a letter signifying the section length between load coils and a number signifying the coil inductance in millihenrys. In voice frequency loading systems, the section lengths generally used are a nominal 6,000-foot (*H* system) and the 3,000-foot (*B* system) although other systems like *D* and *M* with 4,500- and 9,000-foot spacings are still in use in some areas. *H-88* loading on cables for voice frequency service is considered by many telephone companies to be the standard method of loading. The use of the letter-number designation is an improvement over the older designation for loading which was originally known as "heavy," "medium," or "light" loading. Subsequent use of "extra-heavy," "extra-light," "medium-heavy," and "light medium" developed confusion—thus the need for the present system of specifying the type of cable loading.

The letter-number method of designating cable loading is applicable also to phantom group loading where a third communication circuit is derived from the use of two physical circuits called side circuits. The letter *S* (side circuit), and the letter *P* (phantom circuit), are used in certain cases to specify a side and phantom circuit of a phantom group respectively, e.g., *H-88-S* or *H-50-P*. In the example cited, the nominal impedance of the side circuit is 1,100 ohms while that of the phantom circuit is 650 ohms. The nominal impedance values of loaded telephone cables for exchange use will vary from 700 ohms for *H-44* loading to 2,200 ohms when *B-175* loading is used.

Fig. 2 illustrates impedance variations with frequency for commonly used gages of nonloaded cables. The smaller gages of cable have a higher impedance for any given frequency. Fig. 3(A) illustrates impedance variations with frequency for 19-gage cable with *H88* loading commonly designated as *19H88*. The curve illustrated has no irregular changes of amplitude over the range of 200 cycles to 3,400 cycles; therefore, the cable facility which the curve represents has no impedance irregularities. Also shown in Fig. 3 is the average value of the structural return loss for a *19H88* cable.

Fig 3(B) shows the impedance characteristics and the structural return loss curve for a *19B88* facility, while Fig. 4(A) illustrates the impedance curve for a *19D88* facility, and Fig. 4(B) shows the characteristics of a *19M88* facility with a cutoff frequency at approximately 2,800 cycles. Certain areas may use each type of loading system mentioned.

Finer gages of loaded cable will have an impedance curve which is nearly the same as the one for the coarser gages of loaded cable. This holds true for each type of loading. When the impedance curve of nonloaded cables is compared with that of loaded cables, it may be noted that the impedances of the two types of facilities are quite dissimilar and cannot be matched with the use of transformers because of the great differences in slopes of the curves.

Cable impedance irregularities, with any type of loading, will be minimized when the loading coils are placed at exactly the theoretical distance. This presupposes that the entire length of a given cable has the same value of resistance and capacity per unit length.

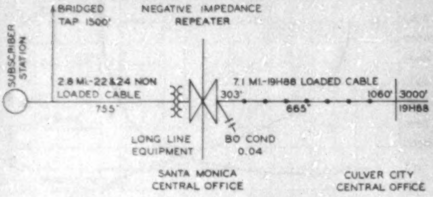

SUBSCRIBER STATION
BRIDGED TAP 1500'
NEGATIVE IMPEDANCE REPEATER
2.8 MI.-22 & 24 NON LOADED CABLE
755'
7.1 MI.-19H88 LOADED CABLE
303'
665'
1060'
3000' 19H88
LONG LINE EQUIPMENT
BO COND 0.04
SANTA MONICA CENTRAL OFFICE
CULVER CITY CENTRAL OFFICE

Fig. 5 (above). Foreign exchange service

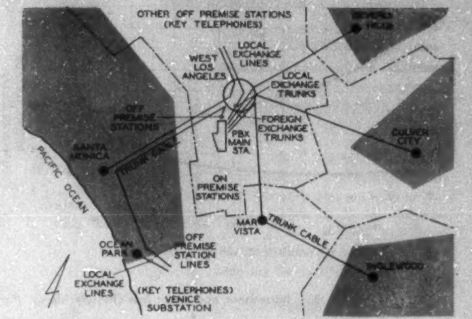

OTHER OFF PREMISE STATIONS (KEY TELEPHONES)
WEST LOS ANGELES
LOCAL EXCHANGE LINES
LOCAL EXCHANGE TRUNKS
OFF PREMISE STATIONS
FOREIGN EXCHANGE TRUNKS
PBX MAIN STA
SANTA MONICA
TRUNK CABLE
PACIFIC OCEAN
CULVER CITY
ON PREMISE STATIONS
MAR VISTA
TRUNK CABLE
OCEAN PARK
OFF PREMISE STATION LINES
INGLEWOOD
LOCAL EXCHANGE LINES
(KEY TELEPHONES) VENICE SUBSTATION

Fig. 6 (right). Typical layout of special service lines

Because of the previously mentioned practical aspects of plant layout, such as the location of rivers, of manholes, or in some cases, of poles, it is not possible to place load coils to conform exactly with the ideal spacing. It has been determined that good return losses will be obtained if the following limits are met:

1. The average loading section length shall be within ±2% of the standard spacing.
2. Each section length shall be within ±2% of the average section.
3. The average absolute value of the individual percentage deviations from the average loading section shall not exceed 0.5%.

The effect of transmission improvement attributed to the use of loading coils reduces the attenuation constant and is usually called "loading gain." For a loading system having three or more load points, effective loading gain in decibels, including the effects of telephone side tone, can be expressed by an empirical term in the form of aN-b. N represents the number of load points, and the constants a and b depend upon the type of loading and the gage of the wire. For transmission design objectives, the 19-gage *H88* loading gain is normally calculated from the term $1.1N - 1.5$, and cable having lengths of 22- and 24-gage from the term $1.5N-1.5$.

Besides taking advantage of loading gains to help meet transmission objectives, it is customary to zone telephone sets according to their transmission efficiency. The higher-efficiency instruments are usually placed at points farthest from the source of d-c power for the telephone transmitter while less efficient sets are zoned in regions near the central office. Considering both the telephone transmitter and receiver, the load loss will be relatively the same for all offices, thus making it possible for the control of overall transmission from subscriber-to-subscriber connections to be dependent upon specific exchange trunk loss. Design objectives for telephone trunks are usually established for a maximum net trunk loss not exceeding 6 db. This net loss includes, among other factors, the loss due to load-spacing irregularities, junction losses, effective trunk facility loss, and repeater gain.

If the sum of the trunk losses, including repeater gain, does not exceed the 6-db value, the transmission objective will be attained and trunk design can be made independent of subscriber loop design. Ideally, if all trunk circuits could be operated at zero loss, the transmission loss on all connections would be independent of the number of circuits used in the connection. However, even with the availability of low-loss circuits, such as those derived from carrier, radio, and microwave systems, the optimum transmission solution is to operate every circuit at some finite loss, which is practical when considering its length and the type of facility used.

When different types of cable facilities are connected together to meet the telephone traffic demands between widely separated exchanges with the minimum transmission loss, certain additional construction limitations are desirable. On through-switched circuits, the loaded cable pairs should terminate in each central office with an end section of one-half the length of a full section between

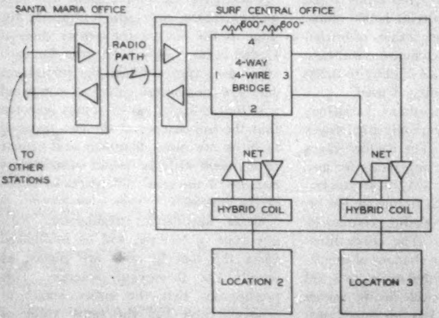

SANTA MARIA OFFICE
SURF CENTRAL OFFICE
600
600
4
RADIO PATH
4-WAY 4-WIRE BRIDGE
1
3
2
NET
NET
HYBRID COIL
HYBRID COIL
TO OTHER STATIONS
LOCATION 2
LOCATION 3

Fig. 7. Private line utilizing a 4-wire 4-way bridge

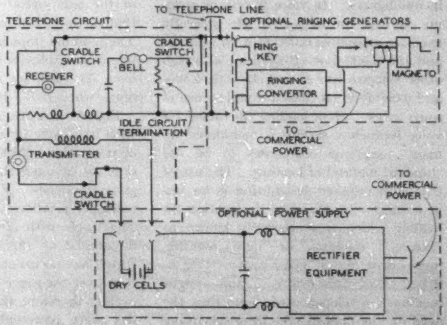

TO TELEPHONE LINE
TELEPHONE CIRCUIT
OPTIONAL RINGING GENERATORS
CRADLE SWITCH
CRADLE SWITCH
BELL
RING KEY
RECEIVER
RINGING CONVERTOR
MAGNETO
IDLE CIRCUIT TERMINATION
TO COMMERCIAL POWER
TRANSMITTER
CRADLE SWITCH
TO COMMERCIAL POWER
OPTIONAL POWER SUPPLY
DRY CELLS
RECTIFIER EQUIPMENT

Fig. 8. Local battery telephone

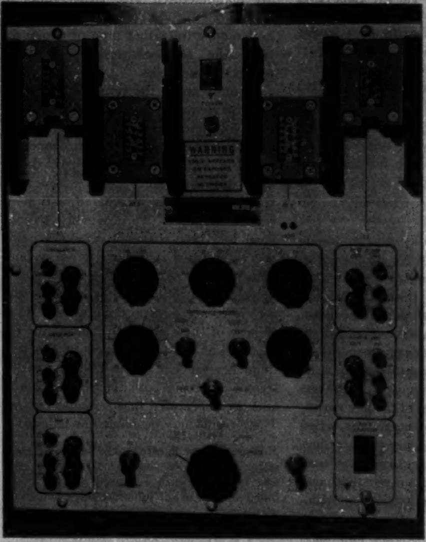


Fig. 9 (left). Repeater test set

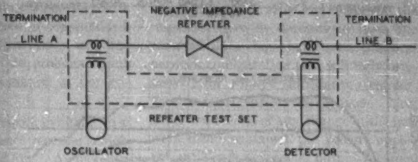


Fig. 10 (above). Line under test

load points. When another loaded cable having the same length of end section is switched to the first cable, a full cable section, properly loaded, will result. However, central offices are not always separated by a distance equal to a whole-number multiple of the loading sections. End section lengths greater than, or less than, half-sections may result. Under these conditions, impedance irregularities will be created and the performance of negative impedance repeaters that may be used on such lines will be adversely affected. Specifically, the usable gain from a negative impedance repeater connected to such a cable pair will depend directly upon how well the actual impedance matches the nominal impedance value of the cable pair.

Many special circuits, including foreign exchange (FX) service, pose requirements for transmission which are not contemplated by normal plant arrangements and cable loading. Fig. 5 illustrates an FX service (Culver City service in Santa Monica) which requires the use of long-line equipment in the Santa Monica impedance repeater to meet the transmission requirements. The existence of a bridge tap, or attached length of unused cable near the telephone set in Santa Monica, causes a transmission loss which must be considered when the required repeater gain is determined. Because of the short-end section of cable, 303 ft (feet), at Santa Monica, additional capacity amounting to 0.04 microfarad is required to build out the cable to an equivalent half-section of 3,000 ft. The irregular loading adjacent to the Culver City Central Office (1,060+3,000=4,060 ft) will cause an impedance irregularity until it is corrected. Fig. 6 illustrates the special service lines used by the Police Department located in the West Los Angeles district. Transmission requirements must be met with a maximum contrast not to exceed 3 db when connected to any combination of trunks accessible by the use of key sets or the use of a private branch exchange board.

Private (Leased) Lines

Quite often large business concerns have offices in different cities or in different states which must be continuously interconnected. An example of a long private line is one used by the Air Force consisting of leased facilities between the missile launching site at Cape Canaveral on the East Coast and the launching site at Vandenberg Air Force base on the West Coast. In many cases, conference circuits are also required at a number of locations at each end of the communication circuit. Loaded cable facilities usually are not used for circuits of great length because of their relatively low velocity of propagation (of the order of 14,000 miles per second). Instead, radio or carrier circuits are usually used in order to take advantage of their characteristically high velocity of wave propagation, thus reducing the problems of electrical echo.

As the time delay in receiving an echo increases, the disturbing effect also increases and becomes very annoying to a speaker. Long lengths of voice frequency loaded cable circuits would have such poor echo characteristics that they normally could not be used without the benefit of echo suppressors for improving voice transmission.

Voice frequency negative-impedance repeater combinations used on practically perfect 2-wire circuits are limited to a gain value of less than 85% of the facility loss with a maximum gain of approximately 12 db. When greater values of repeater gain must be used, separate voice frequency amplifiers for each direction of operation can be connected on a 4-wire basis to two cable pairs assigned to one circuit. By using hybrid coils, these repeaters can also be used on 2-wire circuits. The addition of private line conference circuits connected for 4-wire operation requires the use of 3-way, 4-way, or 6-way bridges depending upon the number of nearby locations to be connected.

A 4-way 4-wire bridge connected to a leased line circuit is illustrated in Fig. 7. To retain an electrical balance of the bridge, the impedance of the network associated with a hybrid coil should match the impedance of the 2-wire circuit connecting the coil to a location. Input-output connections that are not in use will normally be terminated with a resistance equal to the characteristic impedance of that circuit. To avoid upsetting the balance of the circuit connected to the hybrid coil, whenever the telephone receiver is placed on the hook switch, an idle-circuit termination is required. Idle-circuit terminations will simulate, during the idle condition, essentially the same impedance as that presented in the talking condition. Direct current for the telephone transmitter may not be sufficient for proper operation when the connecting wire resistance causes the current to be

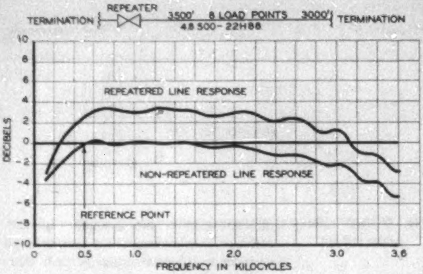


Fig. 11. Gain versus frequency for repeater on a loaded line

reduced to less than 20 milliamperes. When this condition is encountered, use can be made of telephones with a transistor amplifier in the talking path powered from the central office or telephones designed for local battery operation. Fig. 8 illustrates a local battery telephone set with optional connections for ringing and transmitter power.

Repeater Gain Problems

The two basic types of repeaters available to provide transmission gain on telephone lines are the hybrid type and the negative impedance type. Hybrid-type repeaters are usually more costly to place in service because amplification in both directions requires the use of hybrid coils to provide the needed 4-wire connections. When the balancing network has an impedance that exactly matches the 2-wire line which is connected to the hybrid coil, optimum gain from the repeaters can be realized. Adequate matching of the balancing network usually cannot be obtained when impedance irregularities exist on the connecting cable. Under these conditions a spare cable pair from the same cable complement (with a far-end termination) may be connected to the balancing network terminals to obtain optimum gain with no adverse effects on the repeater voice frequency response.

With the negative impedance type of repeater the voice frequency response varies as the impedance of the line varies. Even with this disadvantage, the series type of negative impedance repeater has gained popularity because of its lower cost and its ability to maintain signaling and voice frequency continuity with or without gain. When the active impedance converter fails, the insertion loss may approach 15 db, depending upon the type of repeater. However, the d-c pulsing path will not be disrupted as long as the series-connected repeat coil windings do not fail. Shunt-connected negative admittance repeaters do not have signaling or voice frequencies passing through them; therefore, failure of the active section of the repeater will not completely disrupt a circuit but will merely reduce the gain.

After the required gain of a repeater has been determined, the installing, connecting, and testing of the repeater is usually accomplished with the aid of a repeater test instrument similar to the one shown in Fig. 9. Becuase repeaters are manufactured with either tubes or transistors, the test set must be capable of accommodating either type of design for both the series-connected or shunt-connected repeators.

In addition to the repeater test set a variable frequency oscillator and vacuum-tube voltmeter, serving as a signal source and a detector respectively, are connected as in Fig. 10. This setup is used to determine data for gain versus frequency response curves on repeatered and nonrepeatered lines. To obtain a reference level for these curves a reference point was established at 500 cycles per second. This point corresponds to −40 db on the detector, which in turn is used as the zero reference level on the graphs. With the use of these reference values, line response was plotted throughout the test frequency range of 200 to 4,000 cycles per second to obtain both the nonrepeatered and the repeatered line response curves. These terms may be defined as follows:

Nonrepeatered line response is the attenuation response of the line throughout the test frequency range caused by the combined effects of all inherent impedances as witnessed at the repeater point.

Repeatered line response is the attenuation response of the line after the repeater is operative.

While the curves indicate no "real" values of attenuation, gain, or impedance, they do illustrate the action of the repeater when it is applied to a line at any given test frequency. The vertical distances between the two curves are the true values of gain for the repeatered line.

Fig. 11 shows the decibel gain or loss versus frequency up to 3,600 cycles for a 22-*H88* loaded line with eight load points; with one end section of 3,500 feet and the other end section of 3,000 feet; with compromise terminations of 900 ohms in series with a 2-microfarad con-

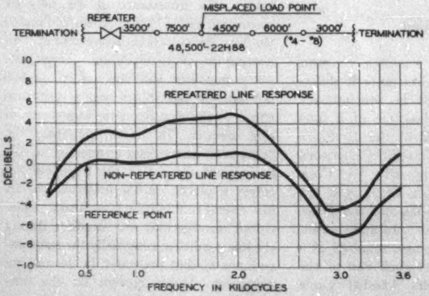


Fig. 12. Curve showing effect of a misplaced load point

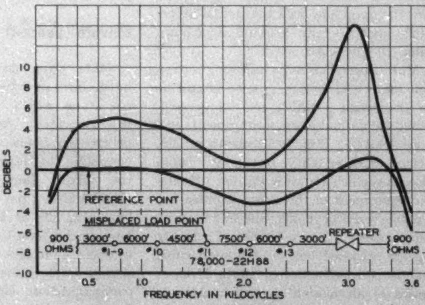


Fig. 13. Curve showing effect of repeater on line with eleventh load point misplaced

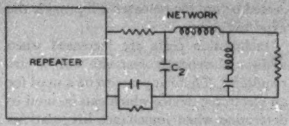


Fig. 14. **General network configuration for repeaters used with loaded lines**

denser. The second curve in the figure shows the gain obtained when the series negative impedance repeater is connected in the line. Note that the shape of the curve with the repeater in use is essentially the same as the curve showing only the line characteristics.

True or average gain between 500 and 3,000 cycles can be obtained from the curves by recording the difference between the amplified curve and the line curve at each peak and valley within the frequency interval, then summing the values in the usual way and dividing by the total number of peaks and valleys considered. The required gain of the repeater shown in Fig. 11 is 3.0 db and the average gain is 3.2 db when nine peaks or valleys are considered. If the same repeater is connected to a line having the second load coil misplaced by 1,500 feet, the repeater response will show a loss of gain for frequencies above 2,200 cycles as illustrated in Fig. 12.

When gains approach losses on a line, there will be oscillation or singing. Such a condition is illustrated in Fig. 13 which shows a high gain at a frequency of approximately 3,000 cycles. The curve shown is for a repeater placed on a line having 13 load points with the third load coil misplaced. Reduction of the gain at high frequencies can be obtained by a change of the values in the network; however, while this may correct the singing condition, the impedance irregularity is still present and may cause transmission problems arising from electrical echo effects. In addition, if nonstandard repeater network adjustments are resorted to, administrative problems may develop in the maintenance of the repeater.

Fig. 14 shows the network available in a series negative impedance repeater which is connected to a loaded line. An increase in the value of the C_2 capacitor will lower the gain for the frequencies between 1,500 and 3,000 cycles. If the same repeater is placed on a cable pair that does not have an impedance irregularity which causes a peak in the response curve at 3,000 cycles the oscillation will cease and the response curve will be similar to that shown in Fig. 11.

Different values of terminations will cause a slight change in the shape of the response curves as shown in Fig. 15(A) and (B) for a mid-section repeater. Fig. 15(A) shows a repeater with a 1,100-ohm noninductive resistor at each end while Fig. 15(B) shows a repeater with a 900-ohm resistor at one end of the line, and at the other end a 900-ohm resistor in series with a 2-microfarad condenser. Normally, more gain can be realized from a repeater with a mid-section installation than from a terminal installation. This is because the highest sum of return loss value is obtained when looking both ways from an intermediate installation as compared to essentially a zero return loss seen in one of the directions of a terminal installation.

Drastic changes occur in the negative impedance repeater response when the repeater is connected to a cable pair with one load point missing. Fig. 16 illustrates the repeater response when it is connected with a 19-*H88* cable facility with one loading coil absent. In most cases, a change in the value of repeater network components will alter the shape of the response curve as determined from the recorded data, but there is no practical network configuration that will fully compensate for a major impedance irregularity that is external to the repeater. Normally the results of any change in the network components will not become apparent until each new set of data is recorded and analyzed. Such a procedure may require an exceptional amount of time when negative impedance repeaters are connected to cable pairs having impedance irregularities. Excessive installation costs will result unless some faster method of testing and recording results of negative impedance repeater operation is used. Such a testing, display, and recording technique is available and will be treated in a companion paper.

Conclusions

Capacity and inductance are the major parameters that affect the value of cable impedance. Variations in capacity may amount to 14% from minimum to maximum for certain cable pairs in paper-insulated lead-covered cables. These variations in capacity have a direct effect on values of structural return loss of a cable since the return loss is dependent on the cable impedance.

Each type of loading system has a different value of impedance; any misplacement of loading coils within the loading system or variations in cable capacity between loading coils will cause impedance irregularities to exist. The proper choice of end sections is important if load-spacing irregularities are to be avoided on switched-through connections.

Negative impedance repeaters connected to cable pairs contain a network

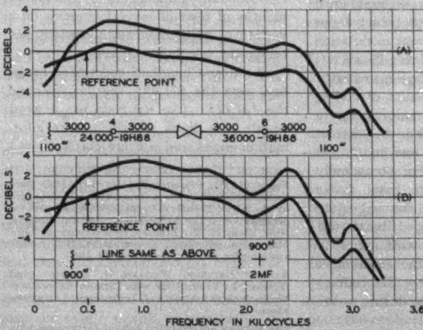


Fig. 15. **Curves showing effects of different terminations**

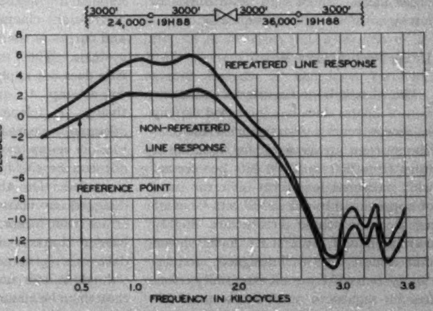


Fig. 16. **Curves showing effect of missing load point**

which can be changed to correspond with the type of cable in use and the amount of gain required. Response curves for a repeater installation are normally obtained by using a repeater test set in conjunction with a signal generator and a detector. Two curves are required, one showing the nonrepeatered line and one showing the repeatered line which indicates the apparent inserted gain. Average repeater gain is obtained between 500 and 3,000 cycles to determine whether it compares favorably with the required gain.

Any changes made in network components will normally require a retest of the repeater in order to determine the magnitude of the repeater gain at any specified frequency. Several test runs may be necessary before a negative impedance repeater is properly installed on a cable pair, and in some cases impedance irregularities within the cable must be corrected before the repeater will provide the expected gain.

Installation costs are increased when additional repeater response curves are required. There appears to be a need for an electronic device which can be used to determine when impedance irregularities exist on a specified cable pair and which can quickly display the proper operations of negative impedance repeaters placed on telephone lines.

The Design of an Error-Free Data Transmission System for Telephone Circuits

BARNEY REIFFEN
NONMEMBER AIEE

WILLIAM G. SCHMIDT
NONMEMBER AIEE

HOWARD L. YUDKIN
NONMEMBER AIEE

RECENT EXPERIMENTAL results verify that easily implemented codes can detect essentially all errors occurring in data sent over toll-grade telephone service. A 2-way communications system uses these codes, first to find and then to request a retransmission of any data in error. The feedback logic is detailed herein, and the buffer required to adapt the system to various data sources is discussed. Extrapolated experimental results indicate an average delivery rate in each direction of approximately 7/8 the modem (modulation system) bit rate with a mean time to error of several hundred years. The philosophy which guided the design of this system can be applied to other media, where high noise bursts or low signal levels occur infrequently.

I. A Preliminary Word

Use of telephone circuits for transmission of digital data has increased rapidly in recent years. A number of modulation systems are available for signaling at rates from 800 to 2,400 bits per second with measured long-term bit error rates of approximately one in 10^5 bits.[1] Most significantly, these errors are highly correlated. Measurements have shown that almost any well-designed modulation system will transmit essentially error-free bit sequences much in excess of 10^5 bits; these error-free periods are terminated by relatively short periods with many errors. Waveforms at the output of the telephone circuit show these error periods are attributable to large energy bursts or circuit dropouts, frequently caused by thunderstorm activity and/or switching transients,[2] and are independent of the modulation system used.

The mean error rate is sensitive to the averaging period, whether a minute, an hour, or a day; averaged over 24 hours, it frequently differs by many orders of magnitude on successive days.

The communications system engineer naturally turns to the information theorist and asks, "Is it possible to employ coding to reduce the effects of bit errors?" The bulk of coding theory work has been applied to memoryless (without error correlation) channels, with binary input and output as used in current modulation systems. The memoryless characteristic does not apply to telephone circuits. An attempt to apply results of coding theory to telephone circuits poses a dilemma. Codes designed to correct random errors, averaging one in 10^5, will not correct the bursts characteristically measured. On the other hand, codes designed to correct for the aggregate errors that occur during a burst will burden the transmission with a loss in data rate (governed by the addition of redundant bits implied by the coding) that may be unacceptable during noiseless periods.

Recently, several workers have devised burst error-correcting schemes that can correct any pattern of errors that lie within a fixed burst length.[3,4] However, during noisy periods, bursts have been observed with many hundreds of bits in error. The design of equipment to correct bursts of length 100 or more is impractical. While the encoder may be relatively simple and straightforward for large classes of codes, the complexity of the required decoding equipment is usually prohibitive.

If error detection is acceptable, in place of detection and correction, then decoding can become comparable to encoding in simplicity. Also, error detection can be achieved with fewer redundant bits than required by a code that can correct the same fraction of errors. Furthermore, as pointed out, if a feedback channel is available, correction can be achieved by utilizing error detection and employing the feedback channel to request the retransmission of those words in which errors have been detected.[5–7]

Motivated by these considerations, Fontaine and Gallager, using Bose-Chaudhuri codes, simulated the communication process on a *709* computer.[8,9] The ability of these codes to detect telephone circuit error patterns was tested against data gathered in a number of months of tests of several modems on a variety of voice bandwidth nonswitched toll-grade telephone circuits.

The Bose-Chaudhuri codes used in the test were the *511,493*, the *255,239*, and the *127,113*. (An *n,k* code has a word length of *n* digits, of which *k* are information digits and *n-k* are check digits.)

Paper 61-101, recommended by the AIEE Data Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 14, 1960; made available for printing November 29, 1960.

BARNEY REIFFEN, WILLIAM G. SCHMIDT, and HOWARD L. YUDKIN are with Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Mass.

The authors acknowledge helpful discussions with Dr. H. Sherman and R. G. Enticknap.

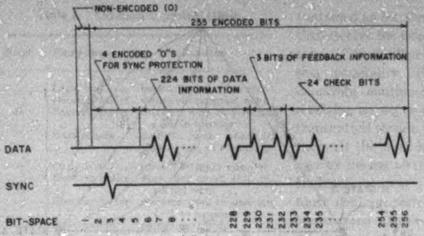


Fig. 1. Format of 256-bit word for feedback communications system using Bell A-1 data system

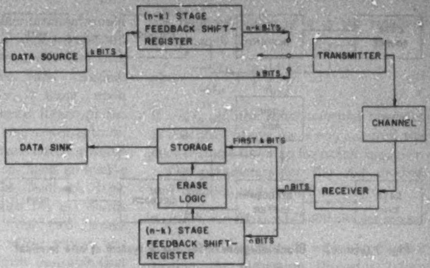


Fig. 2. General encoding-decoding scheme

The first two codes detected *all* error patterns accumulated in the modem tests; the last detected all but one. In each case, approximately one in every 3,000 words was in error.

For the *255,239* code applied to data gathered with the Bell *A-1* data system, an average of approximately one undetected word error in a year may be expected. This estimate was obtained by observing that, with the *255,239* code, an undetected word error can occur only if five or more errors occur in the block, and if the parity checks are satisfied. However, all error patterns are subjected to 16 parity checks. If all error patterns were equally likely, each parity check bit would have a probability of being satisfied equal to one-half. Thus, the probability of an undetected word error is estimated as the product of 2^{-16} and the probability that the data service introduces five or more bit errors in a 255-bit word. A similar estimate of the undetected word error rate of a *255,231* code, calculated on the basis of the same data, yields a figure of approximately 1 per 300 years.

On certain classes of telephone circuits, periods exist when a portion of the circuit is open; this is known as circuit dropout. If it occurs, the signal at the output of a receiver with agc (automatic gain control) of wide dynamic range may consist wholly of high-level random noise. It is useful to calculate the protection that the Bose-Chaudhuri code offers against accepting words consisting wholly of this noise. If it is assumed that all n-bit error patterns are equally likely, an n,k Bose-Chaudhuri code has probability $2^{-(n-k)}$ of accepting an n-bit noise sequence as a transmitted word. Thus, with a 1,300-bit per second modem, we find that *255,239* and *255,231* codes will fail to detect such a noise input, on the average, about once in every 3 hours and once in every month, respectively, of cumulated circuit dropout time.

The foregoing estimates are intended to indicate qualitatively the effectiveness of the class of codes considered for telephone circuits of the type tested. Their verification would be impractical, experimentally. Many years of data would be required to give statistical assurance that the system averaged one undetected word error per year. The important point is: a scheme has been demonstrated with many billions of transmitted bits that can reduce undetected errors, caused by the telephone circuit, to a point where errors, caused by terminal equipment unreliability, become dominant. If, to accomplish error correction, such an error-detection scheme is used in conjunction with a feedback channel, which has an equal degree of error-detection capability, the data transmission error rate is identical to the undetected error rate. It is in this sense that the term "error-free transmission" is used in this paper.

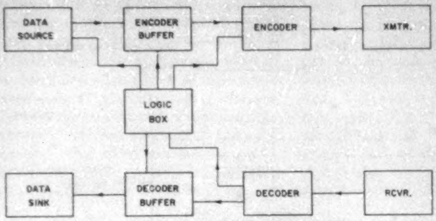
The concept of using a feedback channel to request retransmission of a word in error is not new, but goes back to at least 1951.[5] The novelty of this paper resides in the application of relatively long codes, the implication of almost perfect error detection, and the consequences of such codes in system design.

II. Feedback System Design and Environment

Implementing the concept of error correction, via feedback into a system design, requires the assumption of an environment in which the system will operate, since to build a universal system is beyond reasonable expectations. The authors were motivated by many factors in choosing their assumptions of environment: simplicity without triviality, availability of modems, availability of data on performance of available modems, potential applications, and prejudices—all intertwined. Accordingly, the system design will be described without dwelling, at length, on justifications for the various assumptions. It should be appreciated, however, that other assumptions will give rise to systems which differ in detail, but not in philosophy, from the one described here, which is intended for use on voice bandwidth nonswitched telephone circuits.

The Bell *A-1* data system was selected as the modem.[10] There are now approximately 1,000 of these units in operation, mostly in SAGE (semi-automatic ground environment), and there are more than 1,600 hours of test data, over a variety of telephone media, available for analysis and simulation. The *A-1* serially transmits binary symbols which are detected, using timing derived from a received sync (synchronizing) pulse. The data between sync pulses are called a word. Oscillator stability limits the usable word length to approximately 300 bits. For a Bose-Chaudhuri code, the encoded sequence must be of length 2^m-1, where m is a positive integer. Since at least 16 parity bits (n-$k \geq 16$) are required to obtain an average undetected word error of at least one per year, n should be large to insure that the number of parity bits is not a high fraction of the word length. The superiority of a 24-parity bit system over a 16-bit system in providing protection against periods of high-level random noise was determinative in selecting the *255,231* code.

For ease of implementation, the transmitted word length should be an integral power of 2. Accordingly, the format shown in Fig. 1 was selected. The complete transmitted word consists of 256 bits. One bit is not encoded; it is part of the 5 bits of space bounding the sync for protection. The *A-1* modem must operate without mark bits within 2 time slots of the sync position. Thus, the first 5 time slots in the 256-bit word contain no data. These 5 bits are also



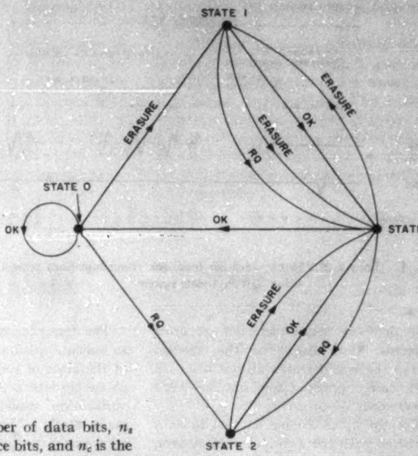


Fig. 3 (above). Block diagram of feedback system at one terminal

Fig. 4 (right). State diagram for a go-back-2 feedback system

inhibited from entering the receiver decoder, 4 receiver-generated space signals being inserted in the place of the encoded transmitter space signals. Hofmann has pointed out that these places are most susceptible to errors.[1] The first 5 bits of the word are called *filler* bits. The next 224 are *data* bits; they contain the actual data being transmitted in that word. The next 3 are *service* bits which convey retransmission requests and information about the state of the channel. The last 24 are the *check* bits of the *255,231* Bose-Chaudhuri code.

The positions of these segments in the word are arranged to facilitate the system implementation. The check bits are placed last in the word. This placement is a consequence of the general encoding and decoding scheme, Fig. 2. The use of shift registers for encoding and decoding this class of codes has been described by Peterson and Meggitt.[11,12] The data source supplies k bits to both the encoding shift-register and to the transmitter. The last n-k bits of the coded word are the contents of the shift-register after the k bits have been encoded. The n bits are transmitted, received at the other terminal, and delivered into the decoding shift register at the receiving terminal, while the first k bits are also inserted into storage. If all stages of the decoding shift-register are not spaces after the n bits have been delivered, then the word in storage is cleared and not delivered to the data sink. The service bits are sent just prior to the check bits where they both may easily be "peeled" from the complete word. This arrangement leads to a minimal time delay between the decision about the correctness of the word and the transmission of its verification. The filler bits are placed at the beginning so that the added bit required to expand the word length to 256 would be used as 1 of the sync protection bits.

The overhead η of the encoding system may be defined as

$$\eta = 1 - \frac{n_d}{n_d + n_s + n_c}$$

where n_d is the number of data bits, n_s is the number of service bits, and n_c is the number of check bits. For the system described,

$$\eta = 1 - \frac{224}{251} \approx 0.11$$

This design assumes that two terminals are communicating with each other over similar channels using the same word format, and that each transmits data and retransmission requests, using identical *255,231* codes. The transmissions are serial and are continuous; i.e., each terminal always sends something. Also, each prints out words in the same order in which they occurred as inputs at the other terminal. Timing around the loop is maintained by designating one terminal as the master with a transmitter time base to serve as the sampling time base for both the transmitter and receiver of the other, known as the slave. The receiver of the master uses the received time base.

It will be shown that the feedback logic requires maintenance of the word synchronization at both terminals. In order to guard against improper synchronization, the *A-1* modem used would have a gated-sync modification incorporated into its operation so that the sync signal may be received only in a small interval of time.

The proposed system is designed for symmetrical data transmission in both directions. A user who requires only 1-way transmission will send dummy data and retransmission instructions over his feedback circuit. This is not unreasonable for the user of telephone facilities, since most toll-grade telephone circuits are 4-wire facilities, 2 wires for sending in each direction. When a 1-way circuit is leased, the channel in the other direction goes unused and can usually be leased for a minor incremental cost. The proposed symmetrical system also eliminates those difficulties attendant in dealing with different signaling systems on the two channels.

The variety of data source characteristics and user requirements are legion. The feedback logic of section II assumes a particularly simple service, but will in fact accommodate other sources which are properly buffered as described in section IV.

III. Feedback Logic

The feedback system concept introduced previously requires a logic which can insure that words transmitted from one terminal to the other are delivered to the data sink, identical in content and order to those the data source delivered to the encoder. Section III presents this logic. An on-demand data source is assumed to deliver words only upon instruction; to wait an infinite amount of time for this instruction; and to deliver data words of the proper length, i.e., 224 bits. Considerations necessary for the logic to accommodate other sources will be given in section IV.

Assume two terminals are communicating with each other, one designated East, and the other West. In the operation of a feedback system, the terminal receiving a word must immediately instruct the other whether or not to retransmit. The command to retransmit a word is

Table I. Go-Back-2 System Logic

State	Component	Function
0....	Logic box	inserts an OK in encoder
	Decoder buffer	prints out received word
	Encoder buffer	delivers word in position $i+1$ to encoder
	Data source	delivers next word to position $i-1$ of encoder
1....	Logic box	inserts an RQ in encoder
	Decoder buffer	does not print out received word
	Encoder buffer	delivers word in position $i-1$ to encoder
	Data source	delays next word
2....	Logic box	inserts an OK in encoder
	Decoder buffer	does not print out received word
	Encoder buffer	delivers word in position $i-1$ to encoder
	Data source	delays next word
3....	Logic box	inserts an OK in encoder
	Decoder buffer	does not print out received word
	Encoder buffer	delivers word in position $i+1$ to encoder
	Data source	delays next word

called an RQ, the alternate instruction is an OK. The last of the 3 service bits, called the verification bit, is used to send the RQ or OK with a mark or space respectively. Each terminal, upon receiving a word, determines if it must send an OK or an RQ, and the time required for this is designated as t_D seconds. The terminal stores its decision until ready to encode its next verification bit.

When a terminal receives an RQ, it retransmits the appropriate word at the earliest opportunity. If a terminal receives an erasure, i.e., a word in which errors have been detected, the assumption is that the erased word contained an RQ. However, if the erased word had really contained an OK, the other terminal would then receive the word an undesired second time. The feedback logic will deal with this possibility. Since each terminal transmits continuously, it must, upon receiving an RQ or erasure, interrupt the transmitted stream of source words to retransmit the word it transmitted q words ago by means of the so-called go-back-q feedback system. The feedback logic must deal properly with the q-1 words transmitted between the two transmissions of the word whose repeat was requested. It is obviously desirable that q be as small as possible; its value is determined by the duration and format of the transmitted word and the 2-way or round-trip delay of the channel.

Let $\tau_{q\,\max}$ be the maximum allowable 2-way delay compatible with a go-back-q system. Suppose both terminals transmit words of n_w-bit length, containing n_c check bits, at a rate of $1/t_B$ bits per second. If the terminals use a go-back-q system on a pair of channels whose 2-way delay is exactly $\tau_{q\,\max}$, each terminal must encode a verification bit immediately after storing the OK or RQ decision. In this event, the relative phasing between transmitted and received words at a terminal is required to be such that the terminal begins transmission of a word at a time $(n_c+1)t_B+t_D$ seconds after the end of the reception of a word. It follows that

$$\tau_{q\,\max}=\{(q-1)n_w-2(n_c+1)\}t_B-2t_D \qquad (1)$$

The derivation of equation 1 is tedious, although straightforward; therefore, it is omitted.

For a given system with 2-way delay τ, the required value of q is the smallest integer such that

$$q\geq 1+\frac{\tau+2t_D}{n_w t_B}+\frac{2(n_c+1)}{n_w} \qquad (2)$$

If $\tau<\tau_{q\,\max}$, the slave terminal operates as described above for $\tau=\tau_{q\,\max}$, and the master stores OK or RQ for a time $\tau_{q\,\max}-\tau$.

The parameters for the *A-1* modem and the assumed word format are

$$n_w=256,\ n_c=24,\ t_B=1/1{,}300 \text{ bits per second}$$

Thus, from equation 1, with a go-back-2 system, the maximum allowable 2-way delay is 158 milliseconds, assuming t_D is negligible. The equivalent 1-way delay of 79 milliseconds is large enough to allow a go-back-2 system to operate between two terminals separated by approximately 790 miles of loaded telephone cable; or 7,900 miles of microwave carrier; or 15,000 miles of hf (high-frequency) radio circuit. This limit is acceptable for many applications, and prompted the design of feedback logic for a go-back-2 system to be described. Modifications needed for a go-back-q system will be indicated. The logic rules are iterative and can best be explained with the aid of Figs. 3 and 4.

Consider each terminal to be as shown in Fig. 3. The block diagram is a simplification designed to help explain the feedback logic. The arrows indicate the direction of information flow.

The logic box represents a centralization of all feedback logical operations, which issue instructions and initiate actions of other system components. The RQ or OK to be encoded in the verification bit of a transmitted word is stored and delivered to the encoder at the proper

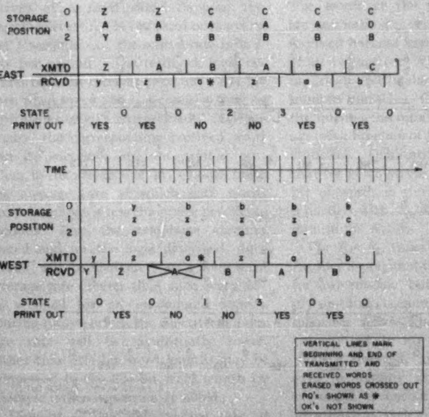


Fig. 5. Operation of go-back-2 feedback system

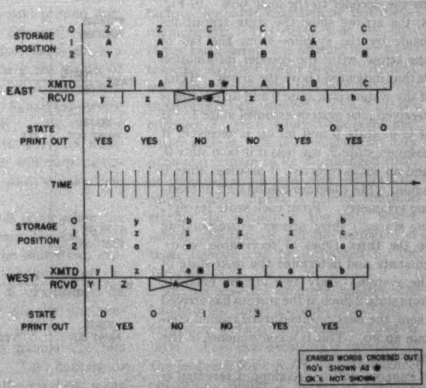


Fig. 6. Variation of Fig. 5

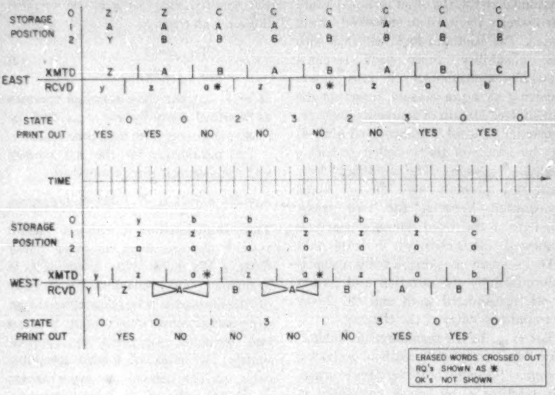


Fig. 7. Variation of Figs. 5 and 6

time by the logic box. The encoder buffer stores three words in positions designated 0, 1 and 2. Encoder buffer storage positions i-1, i, and $i+1$ will be referenced with the understanding that $i-1$, i, and $i+1$ are to be taken modulo 3. The decoder determines if a received word is erased or if it contains an OK or an RQ, and delivers this information to the logic box. The decoder buffer will, upon instructions from the logic box, either deliver a received word to the data sink (print-out) or destroy it.

Let the system have the four states shown in Fig. 4. Upon receiving a word, the system assumes a new state determined by the decoded received word, i.e., by the information about it delivered by the decoder to the logic box. In Fig. 4, the arrows show the state transitions made by the system—from the tail of the arrow, when the decoder information is that indicated above the arrow. Thus, if the system is in state 0 and an erasure is received, the system assumes state 1. If it is in state 0 and an RQ is received, it assumes state 2, whereas if it is in state 0, and an OK is received, it assumes state 0 again. The diagram exhibits an interesting symmetry. From each state (excepting state 2) three arrows, corresponding to the three types of receivable words, emanate and determine the next state of the system. Only two arrows emanate from state 2 since, if the system has arrived there, the next received word cannot contain an RQ. This is a consequence of the rules to be described.

Suppose during decoding, a terminal is transmitting in storage position i. If the word causes the system to assume the state shown in the first column of Table I, the components in the second column, acting upon instructions from the logic box, perform the functions listed in the third.

Note in Fig. 4 that all transitions from states 1 and 2, and none from state 0, go to state 3. Table I shows that when a word causes a transition to state 3, this word is effectively ignored. These rules are demonstrated serially in Figs. 5 through 8, in which both East and West terminals are assumed to start in state 0.

In Fig. 5, West receives an erasure of word *A* and the RQ is decoded by East. In Fig. 6, the RQ is received as an erasure at East. In both cases, each terminal returns to 0 and prints out the previously erased word in proper sequence. If, in either case, the first transmission of word *B* or retransmission of word *z* is erased, no difficulties will arise because of action taken by the system when it is in state 3. Now, if either the retransmission of *A* (see Fig. 7) or *a* (see Fig. 8) is erased, the system cycles again until both words have been correctly received and printed out. If these retransmissions are continually erased, the system persists in cycling until each word has been correctly received and printed out.

It has been assumed throughout that the system is, in fact, error-free. Consider a particular error pattern in which the resultant incorrect word is not erased by the decoder. If the transmitted word contained an OK, or RQ, and the received incorrect word also contains an OK or RQ, the feedback system will not be affected, but an incorrect word will be printed out at one terminal. If, however, the transmitted word contained an OK, or RQ, and the received word contained an RQ, or OK, one terminal will print out two words twice and the other will fail to print them. After that, normal operation is resumed. It should be noted, however, that the encoding and decoding procedure chosen has made improbable the occurrence of this situation.

The rules given in this section for a go-back-2 feedback system generalize in a straightforward way for the go-back-*q* system. In Fig. 9, the state diagram for a go-back-*q* system is shown. Note that state 3 has been replaced by states 3_1, $3_2, \ldots, 3_{q-1}$.

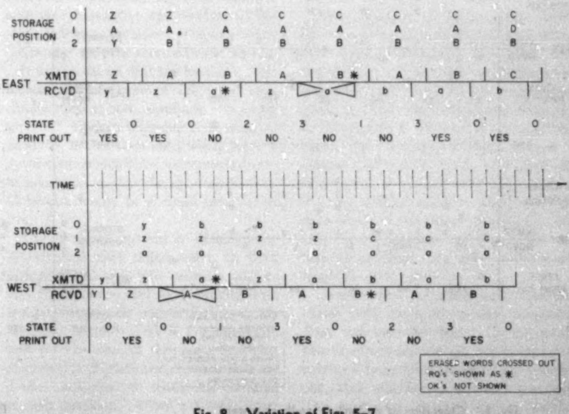


Fig. 8. Variation of Figs. 5–7

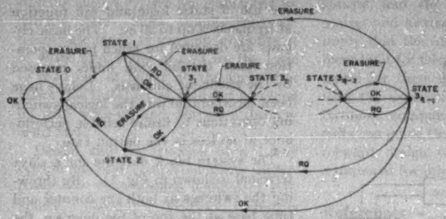


Fig. 9. State diagram for a go-back-q system

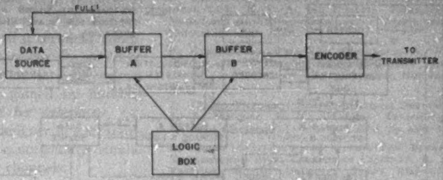


Fig. 10. General input scheme

If the system is in state $3_j (j<q-1)$, it automatically assumes state 3_{j+1} next. In this case, the encoder buffer has $q+1$ storage positions. The operations to be performed in each state are identical with those given for the go-back-2 system if $i-1$ is replaced by $i+1-q$ (modulo $q+1$) and the rules used in every state $3_j (j=1, \ldots, q-1)$ are the same as were used in state 3. The feedback logic described here is not unique. Related logics have been suggested previously.[5–7]

IV. Storage Considerations

The word format and feedback logic discussed in sections II and III assume an on-demand data source, i.e., delivery of 224 data bits whenever requested by the logic box. Not all, if any, data sources are so obliging!

However, if the actual data source is an on-demand type that delivers d-bit data words ($d \neq 224$), it may be converted to a 224-bit word source by a buffer of straightforward design.

A more difficult problem arises if data arrives at a fixed rate. Suppose the source delivers a 224-bit word once every 224 T seconds, i.e., the source rate is $1/T$ bits per second. The feedback logic requires that the receiver print out a 224-bit data word every 256 t_B seconds so long as no erasures or RQ's are received. If either occurs, the corresponding received word and its successor are not printed out. Thus, in the presence of any channel noise the average rate at which data words are printed out is less than once per 256 t_B seconds. For the telephone circuits tested and for the logic described, data words will be delivered at a long-term average rate greater than once every 257 t_B seconds for an on-demand source. During noisy periods the short-term average rate will be significantly lower. Other than 224-bit word lengths may be converted to the 224-bit word length by a simple buffer as previously noted.

Defining R_o as the average rate at which words are printed out for an on-demand source, the words per second would be

$$\frac{1}{257t_B}<R_0<\frac{1}{256t_B}$$

The input rate R_i is equal to $1/224T$ words per second, and the channel rate is $1/t_B$ bits per second.

Since

$$R_i<R_0<\frac{1}{256t_B}$$

then

$$\frac{1}{T}<(7/8)\frac{1}{t_B}$$

Thus, the source bit rate must be less than 7/8 the channel bit rate for the assumptions made.

Since output words are not printed periodically, input words must be stored while a circuit is noisy or open. The number to be stored at one time is a complicated function of R_i/R_0 and the circuit noise statistics. This is a classical queuing problem—once each $224T$ seconds a word arrives at the end of the queue. The words at the front of the queue are serviced, i.e., encoded, transmitted, received without error, and printed out, at an average rate R_o. Successful servicing at any particular time, however, is a random variable. Clearly the length of the queue—the number of words in storage—is a function of R_i/R_0 and the service statistics. Qualitatively, it is obvious that for given service statistics, the number of words in storage will vary monotonically with R_i/R_0, and will become infinite for $R_i>R_0$.

For $R_i<R_0$, times will occur when the encoder is prepared to encode a data word for transmission, but the source can supply nothing because the length of the queue is zero. Since a word must be transmitted to RQ or OK the one previously received, and also to preserve sync, a dummy word having 224 bits of mark is encoded. The dummy is identified as such by the receiver, using the second of the three service bits in the transmitted word, a mark or space in this position meaning a dummy or a legitimate word respectively.

Thus far, it has been assumed that the user desires delivery of all data originating in the source, regardless of delay. But many times this assumption is not valid. For example, the output of an air-defense or traffic-control radar is useful only if it is current. Failure of the channel makes the data obsolete, and retransmission or storage is unnecessary. The proper design of any word storage buffer is intimately connected with the user's requirements. Accordingly, the input system is viewed as in Fig. 10. Buffer A converts the source data into the standard form of 224-bit words, for parallel as well as serial data sources. Furthermore, for a constant rate source, words not as yet transmitted are stored in buffer A. The tandem connection of source and buffer A appears to the rest of the system as an on-demand source delivering words of standard length. Buffer B in Fig. 10 is identical to the buffer in Fig. 3, and requests a new word from A, as described in section III.

Any practical embodiment of buffer A will have a finite word storage, limiting the queue to a certain length. Several factors enter into determining this maximum length. The delay that the customer can tolerate has already been mentioned. Another factor is cost. A third will dominate the first two in many situations of practical interest. This is the occurrence of a gross failure to the communication system. The telephone circuit may open and remain so for a prolonged time, or terminal equipment may fail and require repair. Buffer A cannot continue storing the output of the data source during such periods, but the frequency of these failures can be estimated by the designer.

In the absence of gross failures, the length of the queue has some probability of exceeding a fixed finite length, however large. This probability is a function of the feedback logic R_i/R_0 and the channel statistics. Buffer A should be planned, therefore, with enough storage so that

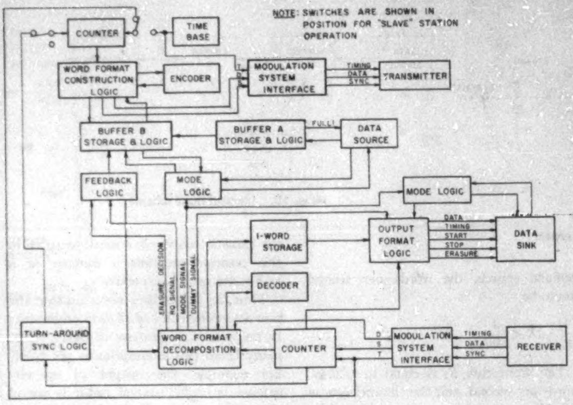

NOTE: SWITCHES ARE SHOWN IN POSITION FOR "SLAVE" STATION OPERATION
COUNTER
TIME BASE
0
WORD FORMAT CONSTRUCTION LOGIC
ENCODER
MODULATION SYSTEM INTERFACE
T
D
S
TIMING
DATA
SYNC
TRANSMITTER
BUFFER B STORAGE & LOGIC
BUFFER A STORAGE & LOGIC
FULL
DATA SOURCE
FEEDBACK LOGIC
MODE LOGIC
MODE LOGIC
1-WORD STORAGE
OUTPUT FORMAT LOGIC
DATA
TIMING
START
STOP
ERASURE
DATA SINK
DECODER
ERASURE DECISION
RQ SIGNAL
MODE SIGNAL
DUMMY SIGNAL
TURN-AROUND SYNC LOGIC
WORD FORMAT DECOMPOSITION LOGIC
COUNTER
D
S
T
MODULATION SYSTEM INTERFACE
TIMING
DATA
SYNC
RECEIVER

Fig. 11. Terminal block diagram

excessive queues are as prevalent as gross failures.

With this point of view, it can be said that such a failure occurs and continues as long as buffer A is full, and the source must be notified thereof. Further data cannot be inserted without loss of that previously inserted. Whether the source stops or continues to insert new data, thereby destroying the old, must be determined by the customer.

The feedback logic, as stated thus far, will not allow data to flow in either direction if one of the circuits fails. For many applications, unidirectional transmission, using error detection without retransmission, may be desirable. A terminal readily recognizes that the circuit connected to its receiver is open, since it will be getting a continuous sequence of erased words. The system may be converted into a 1-way mode whenever the last w words have been erased. The parameter w may be set at the discretion of the user. In the 1-way mode, all words are transmitted once, none repeated.

If one terminal switches to the 1-way mode, the other must be notified that its forward channel is open. This may be accomplished by using a mark or space in the first of the service bits to designate the feedback mode or the 1-way mode respectively. After a terminal has switched to 1-way, it may return to the feedback mode upon receiving a word which has not been erased.

V. System Implementation

As mentioned at the outset, the coding and communications techniques described in this paper place the ultimate reliability of data transmission primarily on the terminal equipment rather than the transmission medium. It is necessary, therefore, that the logical and circuit designs be as reliable as possible. Use of solid-state devices in a controlled environment, both thermally and electrically, should help to achieve the desired results. Special attention must be given to elimination of errors caused by both internal and external noise sources. Not only must the units be shielded and have well-regulated and filtered power supplies, but considerable effort must be expended to insure excellent grounding (without ground loops), confining the noise bandwidth of the transfer lines to a minimum, and maintaining low-impedance levels throughout the equipment rather than merely at the outputs.

To make the system useful with other modems, operation will be permitted at a wide range of bit rates and signal formats. The units now under construction involve the use of static logic and will function at frequencies up to 20 kc. The basic circuits are designed for much faster operation than the bit rate limits imposed by the telephone channels. Delays have been incorporated into the logical switching operations to insure reliable performance at no loss in the output rate.

The system block diagram for a slave terminal is shown in Fig. 11. By throwing the switches at both the counter and time base at the top of the diagram, the block diagram is that of a master terminal.

The basic component of the encoder and decoder is a 24-bit shift-register with certain stages receiving feedback signals through modulo-2 adders. The necessary feedback connections for the *255,231* system shift-register are presented in Fig. 12.

The output system implied by the logic described will deliver at an irregular rate, with gaps occurring where there have been erasures. However, the output words will be in the identical serial order as delivered from the data source unless there has been a channel failure of such duration that the 1-way mode is activated. If the feedback loop of a particular channel has opened, words can be delivered either to the data sink in an erasure format or not at all. An erasure format is one in which a signal is given on a parallel output that the word has had errors detected in its content. Should the data sink desire only error-free data, then the erasure signal can be used to inhibit passage of the word.

To accommodate those data sinks which require delay because of the particular type of recording mechanisms used to store the transmitted data, parallel START and STOP signals bracket the output word. TIMING is also provided in parallel at the output to supply the sink with a time base, needed for sampling and interface conversions.

Transient difficulties associated with

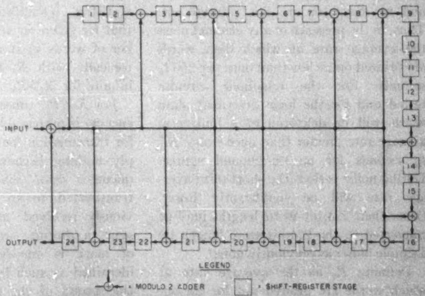

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
INPUT
OUTPUT
LEGEND
MODULO 2 ADDER
SHIFT-REGISTER STAGE

Fig. 12. The 24-stage feedback shift-register for a Bose-Chaudhuri code

starting such a system are eliminated by dummy data. Since simultaneous attempts are made at start-up by buffer A to write into buffer B and by the logic to read out of buffer B, the logic then views this siutation as read-out catching up with write-in, and delivers a dummy word first.

Design and construction of a 1-way system has been completed and is undergoing test to determine equipment reliability. The duplex feedback system is now under construction.

VI. Other Applications

The points of view described in this paper are applicable to any channel, where high noise bursts (i.e., low signal levels or fades) occur at a much slower rate than the channel bit rate with relatively long error-free intervals between bursts or fades. Channels that fall into this general description are: hf, tropospheric, and vhf (very high frequency) scatter and meteor-burst. This does not imply that the identical terminal equipment built for telephone circuits would be useful for these media; it does suggest that the experience gained by building terminal equipment for telephone circuits can guide the designer to a straightforward design of conceptually similar equipment for other channels.

References

1. Measured Error Distributions on the Bell A-1 Facility over Various Media, **E. J. Hofmann.** *Proceedings*, National Electronics Conference, Chicago, Ill., 1960, pp. 37–44.

2. Some Results in the Measurement of Impulse Noise on Several Telephone Circuits, **H. L. Yudkin.** *Ibid.*, pp. 222–31.

3. A Class of Multiple-Error Correcting Binary Codes for Non-Independent Errors, **P. Fire.** *Report RSL-E-2*, Sylvania Electronics Systems Division, Waltham, Mass., Oct. 1959.

4. Recurrent Codes: Easily Mechanized, Burst-Correcting, Binary Codes, **D. W. Hagelbarger.** *Bell System Technical Journal*, New York, N. Y., vol. 38, July 1959, pp. 969–84.

5. Typendrukteleografic over Radioverbindingen TOR, **H. C. A. van Duuren.** *Tijdschrift*, Nederlands Radiogenootschap, Amsterdam, Netherlands, Mar. 1951.

6. Coding for Two-Way Channels, **J. M. Wozencraft, M. Horstein.** Fourth London Symposium on Information Theory, London, England, Aug. 1960.

7. Coded Feedback Communications Systems, **J. J. Metzner, K. C. Morgan.** *Proceedings*, National Electronics Conference, 1960, pp. 250–57.

8. Error Statistics and Coding for Binary Transmission over Telephone Circuits, **A. B. Fontaine, R. G. Gallager.** *Report 25G-0023*, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Mass., 1960.

9. On a Class of Error Correcting Binary Group Codes, **R. C. Bose, D. K. Ray-Chaudhuri.** *Information and Control*, New York, N. Y., vol. 3, no. 1, Mar. 1960, pp. 68–79.

10. SAGE Data Terminals, **R. O. Soffel, E. G. Spack.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 872–79.

11. Encoding and Error-Correction Procedures for the Bose-Chaudhuri Codes, **W. W. Peterson.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-6, no. 4, Sept. 1960, pp. 459–70.

12. Error Correcting Codes for Correcting Bursts of Errors, **J. E. Meggitt.** *IBM Journal of Research and Development*, New York, N. Y., vol. 4, no. 3, July 1960, pp. 329–34.

An Error-Correction Code for Quaternary Data Transmission

E. H. SCHERER
NONMEMBER AIEE

A TREND TOWARD development of nonbinary digital transmission systems has been growing for several years. Much effort has been expended on attempts to optimize digital transmission efficiency, either in terms of speed, reliability, or some other parameter.

Extreme reliability is increasingly demanded by military requirements and by the need for data transmission between data-processing centers serving commercial applications. Channel noise is the chief bar to progress. Because of its source and characteristics, it represents that part of a communication system which is least controllable. Its statistical nature is in sharp contrast to other parts, where an accurate mathematical model can be defined.

This paper presents an error-correction code and a suggested embodiment for encoder and decoder which may be used for improving a quaternary system. No particular application is discussed, however, since applicability rests on a detailed analysis of each individual system and its purpose.

The Quaternary Channel Model

Before the code itself is explained, it is essential to define the channel for which it has been developed. For transmission of information, four discrete signals A through D are available at the transmitter. The decider at the receiving end introduces the additional state E to the alphabet, designating a failure whenever noise in the channel prevents reaching a decision as to which signal has been received. This principle of null-zone decision has been treated in detail for binary transmission.[1]

For quaternary digital transmission with null-zone decision, the channel model is shown in Fig. 1. The lines connecting input and output alphabets represent conditional probabilities. They are designated p_1 through p_5 for transmitted signal A obeying the condition

$$\sum_{j=1}^{5} p_j = 1 \tag{1}$$

where p_1 is the probability of a correctly received signal; p_2, p_3, and p_4 are the error probabilities; and p_5 is the failure probability, when the transmitted signal cannot be recognized and the decider declares a failure.

It is assumed that the probability distribution p_1 through p_5, shown for signal A, applies identically to B, C, and D, the other input signals. In other words, a quaternary symmetric channel is assumed. The information rate of this channel in bits per digit is determined by the difference between entropy and equivocation.[2]

$$R = H(x) - H_y(x) \left[\frac{\text{bits}}{\text{digit}}\right] \tag{2}$$

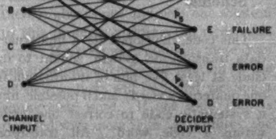


Fig. 1. Channel model for quaternary digital transmission with null-zone decision

---

Paper 61-170, recommended by the AIEE Communication Theory Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 17, 1959; made available for printing December 27, 1960.

E. H. Scherer is with IBM Laboratories, Boeblingen, Wuertt, Germany.

where x refers to the input and y to the output of the channel model. The maximalized rate is equal to the channel capacity C in bits per digit

$$C=[H(x)-H_y(x)]_{\max}\left[\frac{\text{bits}}{\text{digit}}\right] \tag{3}$$

The entropy of one digit is defined by

$$H(x)=-\sum_{i=1}^{4} p_i \log_2 p_i \tag{4}$$

where

$$\sum_{i=1}^{4} p_i=1 \tag{5}$$

Index i refers to the input alphabet A through D. Assuming a random source, the entropy obtained is

$$H(x)=-4\left[\frac{1}{4}\log_2\frac{1}{4}\right]=2\left[\frac{\text{bits}}{\text{digit}}\right] \tag{6}$$

The equivocation is defined by

$$\begin{aligned}H_y(x)&=-\sum_{i=1}^{4}\sum_{j=1}^{5} p(i,j)\log_2 p_j\\&=-\sum_{i=1}^{4}\sum_{j=1}^{5} p(i)p_j\log_2 p_j\end{aligned} \tag{7}$$

When assuming equal probability distribution for errors and failures for all four transmission signals and a random source, the equivocation for this specialized model is

$$H_y(x)=-[p_1\log_2 p_1+p_2\log_2 p_2+p_3\log_2 p_3+p_4\log_2 p_4+p_5\log_2 p_5] \tag{8}$$

If, furthermore, all error probabilities are assumed identical, i.e.,

$$p_2=p_3=p_4=p_E \tag{9}$$

the information rate of this channel is

$$R=2+[p_1\log_2 p_1+3p_E\log_2 p_E+p_5\log_2 p_5]\left[\frac{\text{bits}}{\text{digit}}\right] \tag{10}$$

This equation defines the functional relationship between error and failure probabilities and the information rate of the quaternary channel. It may serve as one starting point for investigating the influence of null-zone decision on the rate without coding.

The Quaternary Code and Encoding Scheme

This code, developed for a quaternary data transmission system, sought to include at least five bits of information in one quaternary word; and if possible, find a systematic code to correct one error within a code word. All transition probabilities p_E were assumed to be equal for all four transmission signals A through D, as would be the case in a system using four mutually orthogonal waveforms for transmission. Since one quaternary digit is equivalent to two bits of information, the number of information-carrying quaternary digits m must be at least three in compliance with the foregoing requirement.

A systematic code can be visualized in terms of Hamming's geometric model of message space.[3] It can be stated that the minimum distance D between any two code words must be at least three since

$$D\geqq 1+2e+f\geqq 1+2(1)=3 \tag{11}$$

where e is the number of errors and f the number of failures that can be simultaneously corrected within one word. Equation 11 shows that error correction can be traded for failure correction on a 1-to-2 basis since this equation is satisfied to $f=2$, $e=0$, or for $f=0$, $e=1$. From another viewpoint, it becomes clear that with a constant amount of redundancy, more failures than errors can be corrected because the location of a failure is known. This is of importance for deciding on the threshold adjustment of the null-zone decider at the receiving end of a system. A code will achieve its full improvement only when the null-zone adjustment of the decider is matched to its constraints.

Hamming's upper bound on the number of digits in an e-error-correcting code with a frame length of n digits, when generalized to an alphabet of a letters, is defined by the inequality

$$a^k\geqq\sum_{i=0}^{e}\binom{n}{i}(a-1)^i \tag{12}$$

where $k=n-m$ is the number of check digits included in a word.[3]

For a quaternary single error-correcting code, $a=4$ and $e=1$, and inequality 12 becomes

$$4^k\geqq 1+3n \tag{13}$$

Hence, this bound can be reached with $k=2$ and $n=5$.

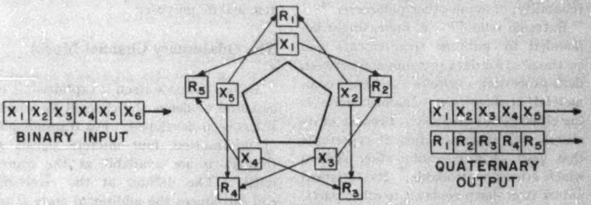
An encoding scheme for such a code was found and it, therefore, may be concluded that this code is optimum in the sense that no better 5-digit quaternary single error-correcting code exists.

A complete set of code words, i.e., the code book, consists of $a^m=4^3=64$ quaternary 5-digit words out of $a^n=4^5=1{,}024$ possible words in the message space. Each code word contains 40% redundancy (two quaternary check digits) relative to the total number of five digits in a code word.

Since a quaternary digit is equivalent to two binary digits, encoding and decoding of quaternary code words can be performed by representing them in binary language and using binary logic.

The binary representation, however, with $a=2$, $n=10$, $k=4$ equals Hamming's upper bound neither for single ($e=1$) nor for double ($e=2$) binary error correction. Therefore, when subjecting the code to disturbances in the channel during transmission, the binary representation is not equivalent to the quaternary code and must not be used unless rather fictitious assumptions are made about the expected error pattern in the corresponding binary transmission channel.

The usefulness of the binary representation will be demonstrated by describing a simple encoder which is capable of generating a complete set of code words. Since three information-carrying quaternary digits represent six binary digits, but



EXAMPLE

DIGIT	X_1	X_2	X_3	X_4	X_5	X_6	R_1	R_2	R_3	R_4	R_5
BINARY INPUT	1	0	1	1	0	0	/	/	/	/	/
QUATERNARY OUTPUT	1	0	1	1	0	/	0	0	1	1	0

DIGIT X_6	0	1
PARITY CHECK	EVEN	ODD

Fig. 2. Binary-to-quaternary encoding scheme

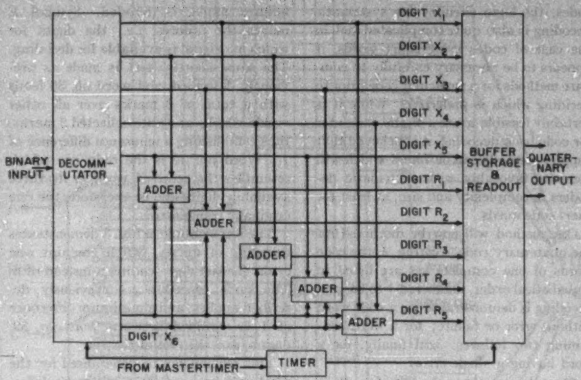


Fig. 3. Schematic diagram of encoder

a complete code word contains five quaternary digits (equivalent to ten binary digits), four binary digits (equivalent to two quaternary digits) must be generated within the encoder. The binary-to-quaternary encoding scheme is shown in Fig. 2. Each input frame contains six binary digits X_1 through X_6 as shown on the left side of the figure. To determine the redundant digits R_1 through R_5, any five of the six input digits may be visualized at the corners of a pentagon. Digits R_1 through R_5 may then be computed by performing five parity checks, as indicated by arrows. The sixth input, digit X_6, is used to determine whether odd or even parity checks are performed for binary to quaternary encoding of this frame

In mathematical, terms the modulo-2 additions performed in the encoder are

$$\begin{aligned}R_1&=X_2+X_5+X_6\\R_2&=X_1+X_3+X_6\\R_3&=X_2+X_4+X_6\\R_4&=X_3+X_5+X_6\\R_5&=X_1+X_4+X_6 \qquad (14)\end{aligned}$$

An encoding example is given in Fig. 2. After the redundant binary digits are determined, the total of ten binary digits are digrammed into pairs for readout, i.e., each pair of digits, one at each corner, is from now on to be considered as the information content of one quaternary digit. Readout could, for instance, be provided at two parallel outputs as indicated on the right side of the figure. Such an arrangement should be considered as the binary representation of the quaternary code word. Whenever the encoder triggers the signal generators in a system directly, it will be more desirable to have four outputs A through D from the encoder releasing a trigger pulse on one, and only one, output at a time in accordance with any binary-to-quaternary conversion scheme as, for instance,

$$\begin{matrix}0\\0\end{matrix}\rightarrow\text{A}\quad\begin{matrix}0\\1\end{matrix}\rightarrow\text{B}\quad\begin{matrix}1\\0\end{matrix}\rightarrow\text{C}\quad\begin{matrix}1\\1\end{matrix}\rightarrow\text{D}$$

This encoding scheme, when applied to all 2^6 combinations of six binary input digits, yields one complete set of 64 quaternary code words at the output.

For encoding, a shifter register and modulo-2 adder could be used in an arrangement similar to the one described by Green and San Soucie for binary codes.[4]

A different approach is shown in Fig. 3 where the parity checks are performed simultaneously in five modulo-2 adders as the binary digits are read in. This encoder has the advantage of being capable of accepting the input digits, either in time sequence as shown, or simultaneously, by switching the input timing circuit to a different mode. The function of the decommutator is to distribute input digits X_1 through X_5 into five parallel lines, which are connected to the buffer storage, and to supply the adders with the information of X_6. Two lines are paralleled into five modulo-2 adders, one for $X_6=0$ and the other for $X_6=1$, to set the adders for even or odd parity checks. The direct connections between decommutator and buffer storage have tappings which are connected to the adders according to the encoding scheme shown in Fig. 2. The adder outputs represent digits R_1 through R_5 and are also connected to the buffer storage. In spite of the fact that R_1 through R_5 could be stored in the adders until they are set or reset, the redundant digits have to be forwarded to the buffer storage, prior to readout, to clear the adders for the next incoming frame.

For a readout sequence as shown in Fig. 2, the encoding delay is a full frame length, since X_2, X_5, and X_6 are required in order to determine R_1. If for some reason the readout sequence is of no importance, the encoder can be built for a minimum encoding delay, with a time interval equivalent to four binary digits. Then X_1 will be used to set all adders for odd or even parity checks. After three more digits, X_2, X_3, and X_4, are read in, the first redundant digit R_3 can be determined. The output sequence will start, therefore, with pair X_3, R_3 at parallel outputs, while X_5 is read in simultaneously.

The timing signals for decommutation, storage, and readout are derived from a timer, which in turn is controlled from the master timer of the terminal.

The development of this code preceded another research effort on the theory and construction of nonbinary multiple error- and failure-correcting codes.[5] This theory appears to contain most of the codes already published, including the present one, as special cases. As a result of this work, it has been established that there are 16 disjoint sets of 64 code words completely filling the message space containing $a^n=4^5=1{,}024$ words, which are obtained by applying all possible combinations of cyclical permutations to any two columns of the listed code words of one set.

The encoder described in this section can be used to produce any of the 16 disjoint sets by properly reversing binary digits in the storage and readout device.

Decoding

There are two basic possibilities for decoding a systematic code, known as systematic and code-book decoding, respectively.

Systematic decoding consists of finding values for X_1 through X_6 and R_1 through R_5 to satisfy the modulo-2 set of equations in 14 of the preceding section. If the constraints of the code are not exceeded, there is always one and only one solution to this problem yielding the right code word. If a code word contains more than one error, one and only one solution for the set of equations will be found yielding a wrong code word. If one code word is received with more than two failures, the solution is ambiguous but includes the right answer. If, therefore, for more than two failures, n sets of values for X_1

through X_5 and R_1 through R_2 satisfy the set of equations and one out of n is selected at random, the probability that this was the originally transmitted word is

$$p_i = \frac{1}{n} \tag{15}$$

It was previously mentioned that systematic codes have a higher failure than error-correction capability. The fact that a residual probability exists for errorless transmission beyond the defined failure constraints further supports the use of a null-zone decider.

Code-book decoding is applicable both to systematic codes and message codes. The function of a code-book decoder is to compare the incoming word with all possible code words of one complete set used for transmission, the set being stored in some form in the decoder. In case of a noiseless channel and faultless equipment, each read-in word will be identical to one of the stored words. If, however, the read-in word is mutilated because of a noisy channel, there will be only one stored word which it closely resembles, provided the failure constraints are not exceeded. If the process of decoding is visualized in the message space, it can be understood as finding the stored word from which the read-in word is at minimum distance.

Code-book decoding is considered a complex and costly method, particularly for long code words, as it necessitates storage of the complete code book. For nonbinary error- and failure-correction codes, the logic circuitry for systematic decoding is also quite complicated and, in the case of codes using short words, it appears to be necessary carefully to compare methods for a particular code prior to deciding which is preferred. While it is certainly feasible to use a digital computer for code-book decoding, it is believed that simpler methods can be found which will compare favorably with systematic decoders in complexity and size, at least for short code words.

One method will now be discussed for the quaternary code. In Fig. 4, the code words of one complete set are listed in alphabetical order, numbered 1 to 64, and decoding is demonstrated for a code word without error or failure; for a word containing two failures; and finally, for a word having a single error.

The column next to the list of words is headed by ADCCD, which serves as an example for one correctly received word. It is compared with the complete code book on a digit-by-digit basis, and a mark is noted for each identity on the respective line. Word no. 7, for example, is identical with received word ADCCD only in its first, A, and third, C, digit. This is indicated by two marks in the proper locations of line 7. By summing all marks, the sums being shown at right of the pattern of marks, a 5 is obtained for word no. 15, which is identical with the received word. All other words have sums of 0, 1, or 2, and there is a minimum difference of 3 between the right word and all the others.

In the next column, a word with two failures BEBEC is decoded. Symbol *E* marks the failures, i.e., the digits for which no signal is available for decoding. The same identity test is made as previously described, and word no. 30 leads with a total of 3 marks over all other words which, at most, collected 2 marks, thus establishing a minimum difference of 1. Word no. 30 is the only one which resembles the received word in its three availabile digits and is, therefore, the one originally transmitted.

The last example in Fig. 4 demonstrates decoding of CBCDA, which contains one error, the last digit reading A instead of D. The same procedure as previously described applies, and a minimum difference of 1 is obtained between word no. 39, CBCDD, and the rest of the set.

This scheme can be generalized for the whole set because these minimum differences will occur for any failure or error in any word of the set. It, therefore, becomes evident that the code-book decoder based on this scheme must perform:

1. Read-in on a word-by-word basis.
2. Decoding. Comparison is made of each read-in word with the a priori stored set of words on a digit-by-digit basis and is followed by summation of identity criteria (marks, charges, voltage steps, etc.) for each word.
3. Interrogation and translation. Sixty-four outputs or sums have to be interrogated to determine the location of the highest sum and its corresponding word number. Translation of the word number into its binary equivalent has to be provided.
4. Storage and readout. The binary equivalent of the decoded quaternary word will be stored until the readout operation, in time sequence or parallel, takes place.

A functional display of these main items is given in Fig. 5 with special emphasis on read-in, decommutation, and decoding.

The decoder has four separate inputs, A through D, over which trigger pulses are accepted from a decider in time sequence, one for each digit of a word. Each input is connected to five lines in time sequence by means of an electronic switch as shown. These switches are operated simultaneously by the decoder timing circuit. They decommutate the input pulse sequence into 20 lines—A1 through A5, B1 through B5, C1 through C5, and D1 through D5. If, for instance, word no. 2, AABCB, is received, the first trigger pulse is received at input A. Since this is the first digit of the word, all the switches are in the top position and only line A1 receives the pulse.

After the first digit is read in, the switches move to A2, B2, C2, and D2,

	Code Word	Correct ADCCD		2 failures BEBEC		1 error CBCDA	
1	AAAAA	+	1		0	+	1
2	AABCB	+ +	2	+	1		0
3	AACDC	+ +	2	+	1	+ +	2
4	AADBD	+ +	2		0		0
5	ABACC	+ +	2	+	1	+	1
6	ABBAD	+ +	2	+	1	+	1
7	ABCBA	+ +	2		0	+ + +	3
8	ABDDB	+	1		0	+ +	2
9	ACADD	+ +	2		0	+	1
10	ACBBC	+	1	+ +	2		0
11	ACCAB	+ +	2		0	+	1
12	ACDCA	+ +	2		0	+	1
13	ADABB	+ +	2		0		0
14	ADBDA	+ +	2	+	1	+ +	2
15	ADCCD	+ + + + +	$\boxed{5}$		0	+	1
16	ADDAC	+ +	2	+	1		0
17	BAABC		0	+ +	2		0
18	BABDD	+	1	+ +	2	+	1
19	BACCA	+ +	2	+	1	+ +	2
20	BADAB		0	+	1		0
21	BBADA		0	+	1	+ + +	3
22	BBBBB		0	+ +	2	+	1
23	BBCAC	+	1	+ +	2	+ +	2
24	BBDCD	+ +	2	+	1	+	1
25	BCACB	+	1	+	1		0
26	BCBAA		0	+ +	2	+	1
27	BCCBD	+ +	2	+	1	+	1
28	BCDDC		0	+ +	2	+	1
29	BDAAD	+ +	2	+	1		0
30	BDBCC	+ +	2	+ + +	$\boxed{3}$		0
31	BDCDB	+ +	2	+	1	+ +	2
32	BDDBA	+	1	+	1	+	1
33	CAACD	+ +	2		0	+	1
34	CABAC		0	+ +	2	+	1
35	CACBB	+	1		0	+ +	2
36	CADDA		0		0	+ + +	3
37	CBAAB		0		0	+ +	2
38	CBBCA	+	1	+	1	+ + +	3
39	CBCDD	+ +	2		0	+ + + +	$\boxed{4}$
40	CBDBC		0	+	1	+ +	2
41	CCABA		0		0	+ +	2
42	CCBDB		0	+	1	+ +	2
43	CCCCC	+ +	2	+	1	+ +	2
44	CCDAD	+	1		0	+	1
45	CDADC	+	1	+	1	+ +	2
46	CDBBD	+ +	2	+	1	+	1
47	CDCAA	+ +	2		0	+ + +	3
48	CDDCB	+ +	2		0	+	1
49	DAADB		0		0	+	1
50	DABBA		0	+	1	+	1
51	DACAD	+ +	2		0	+	1
52	DADCC	+	1	+	1		0
53	DBABD	+	1		0	+	1
54	DBBDC		0	+ +	2	+ +	2
55	DBCCB	+ +	2		0	+ +	2
56	DBDAA		0		0	+ +	2
57	DCAAC		0	+	1		0
58	DCBCD	+ +	2	+	1		0
59	DCCDA	+	1		0	+ + +	3
60	DCDBB		0		0		0
61	DDACA	+ +	2		0	+	1
62	DDBAB	+	1	+	1		0
63	DDCBC	+ +	2	+	1	+	1
64	DDDDD	+ +	2		0	+	1

☐ Maximum Counts Indicate Transmitted Word

Fig. 4. Decoding for correct and distorted words

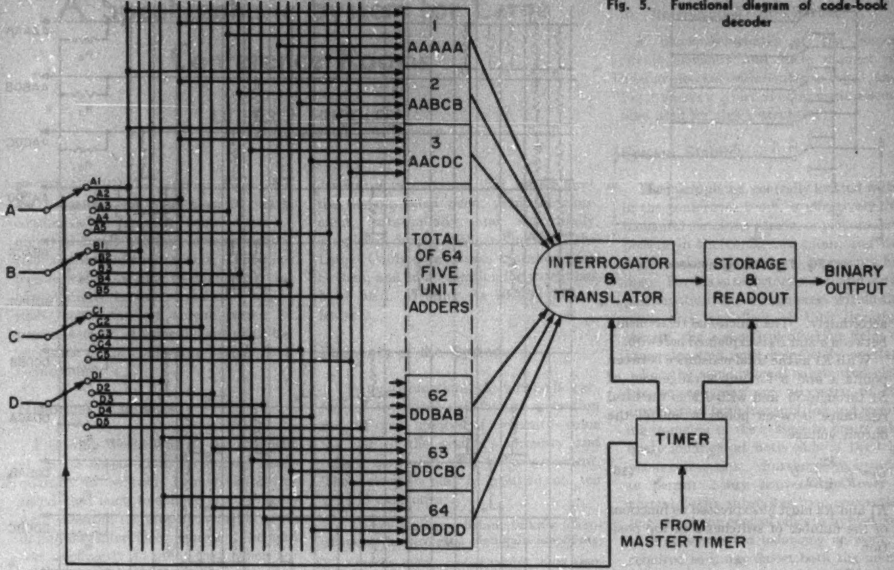

1
AAAAA
2
AABCB
3
AACDC
TOTAL OF 64 FIVE UNIT ADDERS
62
DDBAB
63
DDCBC
64
DDDDD
A
A1
A2
A3
A4
A5
B
B1
B2
B3
B4
B5
C
C1
C2
C3
C4
C5
D
D1
D2
D3
D4
D5
INTERROGATOR & TRANSLATOR
STORAGE & READOUT
BINARY OUTPUT
TIMER
FROM MASTER TIMER

Fig. 5. Functional diagram of code-book decoder

respectively. Another A is read in, and A2 receives this trigger pulse. Then the switches move to the third position, and a trigger pulse is received at input B. Line B3 receives this pulse. The remaining two digits are received by lines C4 and B5. For each of the 64 words of the set, a different combination of five lines will receive the trigger pulses. If, therefore, a wire matrix, as shown in Fig. 5, connects those 20 lines with all adders according to the set of 64 words, all 64 adders will yield the sum of identity criteria (trigger pulses in this example). Word AABCB is decommutated into lines A1, A2, B3, C4, and B5, and the five inputs of adder no. 2 are connected to them. To provide further examples, the input connections of adders no. 1, 3, 63, and 64 are also shown.

The required a priori storage of the complete set of words for the purpose of comparing it with each incoming word is thus provided by the wire matrix. Once a word is read in, it is also decoded automatically because the trigger pulses pass the wire matrix on their way to the adders. The remaining functions are to locate the adder with the highest sum of identity criteria, to translate the respective word number into its binary equivalent, and to provide storage and readout.

A schematic block diagram of the suggested code-book decoder is shown in Fig. 6. There the 20 inputs and 64 outputs of the decoding matrix are indicated. Interrogation and translation are performed in a combined circuit which also contains a diode matrix to translate each of the 64 inputs into its binary equivalent, which is forwarded to the storage and readout circuit by means of six parallel lines.

The basic adding network, suggested for adding identity criteria, is shown in Fig. 7. Five parallel resistors R in series with resistor R_0 are arranged in such a way that each resistor R may be switched independently, either to a supply voltage U or to ground. The purpose of this network is to make U_0 a linear function of the number of switches a, connecting resistors R to the supply voltage. The switches will be operated by those five trigger pulses representing one quaternary word. In this circuit, a therefore can assume any integer between 0 and 5, and U_0 must assume different voltage levels

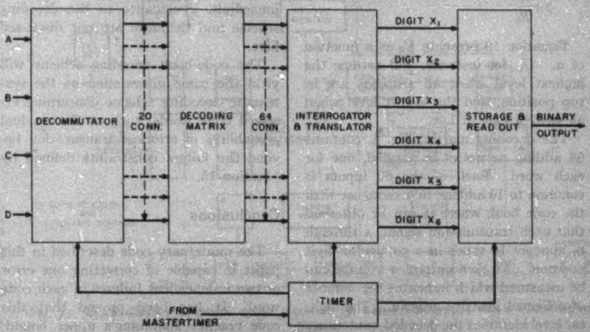

A
B
C
D
DECOMMUTATOR
20 CONN.
DECODING MATRIX
64 CONN
INTERROGATOR & TRANSLATOR
DIGIT X1
DIGIT X2
DIGIT X3
DIGIT X4
DIGIT X5
DIGIT X6
STORAGE & READ OUT
BINARY OUTPUT
TIMER
FROM MASTERTIMER

Fig. 6. Schematic diagram of code-book decoder

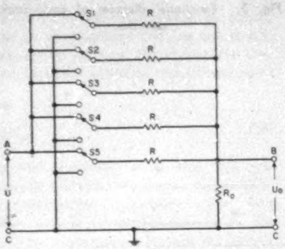


Fig. 7. Adding network

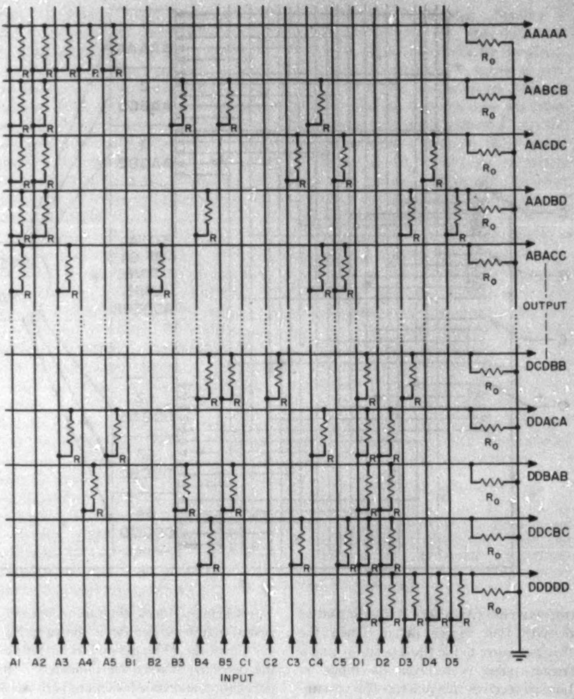


Fig. 8. Decoding matrix

accordingly. The functional relationship between a and U_0 is explained herewith.

With $R1$ as the total resistance between points A and B for any arrangement of $S1$ through $S5$, and with $R2$ as the total resistance between points B and C, the output voltage

$$U_0=\frac{R2}{R1+R2}\,U \qquad (16)$$

$R1$ and $R2$ must be expressed as functions of the number of switches a in top position

$$R1=\frac{R}{a} \qquad (17)$$

$R2$ is the parallel combination of R_0 and $R/(5-a)$

$$R2=\frac{R_0R}{R_0(5-a)+R} \qquad (18)$$

Equations 17 and 18, substituted in 16, yield U_0:

$$U_0=\frac{R_0R}{R_0(5-a)+R}\,\frac{1}{\dfrac{R}{a}+\dfrac{R_0R}{R_0(5-a)+R}}\,U$$

$$U_0=\frac{1}{5+\dfrac{R}{R_0}}\,aU=ka \qquad (19)$$

Equation 19 presents U_0 as a function of a. U_0, for instance, will assume the highest level when all switches are in top position, and the lowest level when all switches are at low.

The decoding matrix, Fig. 8, contains 64 adding networks in parallel, one for each word. Each of the 20 inputs is common to 16 adders, in accordance with the code book where it can be observed that each transmission signal, A through D, appears 16 times in a particular digit position. At each output, a voltage can be measured which indicates the number of collected identity criteria. The highest level identifies the decoded quaternary word.

The remaining functions of the code-book decoder, interrogation of the decoding matrix outputs, translation of the selected output into its binary equivalent, and storage and readout may be handled in many ways. They are not of immediate significance to the decoding scheme and therefore are not discussed here.

The code-book decoding scheme will yield the same information as the systematic decoding scheme concerning received code words, such as the residual probability of errorless transmission beyond the failure constraints defined by equation 15.

Conclusions

The quaternary code described in this paper is capable of correcting one error or two independent failures in each code word. It has been proved that this code reaches Hamming's upper bound. Therefore, no other quaternary code can be found which has equal word length and identical constraints requiring less redundancy.

The described encoding and decoding schemes may be implemented with ease, employing presently available hardware.

References

1. Improvement of Binary Transmission by Null-Zone Reception, F. J. Bloom, S. S. L. Chang, B. Harris, A. Hauptschein, K. C. Morgan. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 45, no. 7, July 1957, pp. 963–75.

2. The Mathematical Theory of Communications (book), C. L. Shannon, W. Weaver. University of Illinois Press, Urbana, Ill., 1949.

3. Error Detecting and Error Correcting Codes, R. W. Hamming. *Bell System Technical Journal*, New York, N. Y., vol. 26, no. 2, Apr. 1950, pp. 147–60.

4. An Error-Correcting Encoder and Decoder of High Efficiency, J. H. Green, Jr., R. L. San Soucie. *Proceedings*, Institute of Radio Engineers, vol. 46, no. 10, Oct. 1958, pp. 1741–44.

5. Results of a Geometric Approach to the Theory and Construction of Non-binary, Multiple Error and Failure Correcting Codes, B. M. Dwork, R. M. Heller. *Convention Record*, Institute of Radio Engineers, vol. 7, pt. 4, Mar. 23–26, 1959, pp. 123–29.

A Speakerphone System for Large Conference Rooms

LUTHER R. HUGGLER
NONMEMBER AIEE

EXECUTIVES of various large corporations, for more than 20 years, continued to request some type of universal hands-free telephone service for their sizable conference rooms. Their objective was to be able to hold important consultations between conferees within such rooms and an absent person or persons located at some distant place but accessible to the conference by telephone connections.

Historical Highlights

Following World War II, the president of an oil corporation, having its headquarters in western Pennsylvania, insisted that such telephone arrangements were essential for proper operation of its day-to-day interoffice business. The need was apparently so great that, when advised by the telephone company that satisfactory arrangements for such service were not available, the oil corporation engaged an electronic technician to devise such a facility.

Assembled under the oil corporation's auspices after several years of effort, this system then was offered for sale to the telephone company. Investigation by communications engineers disclosed, however, that this system's stability was inadequate for use with the general exchange and toll message network of the nation's telephone companies. Also, the effects of conference room noise and reverberation almost completely destroyed the intelligibility of speech signals received by the distant party.

Under the circumstances and after considering advancements made in communications and electronics during the interim, a project was undertaken by the Bell Telephone Company of Pennsylvania, with advice and assistance from the Bell Telephone Laboratories, to develop a new system that would meet the demonstrated need. This development, substantially and successfully completed during 1958, is known as the Large Conference Room Speakerphone System, and an application for patent has been filed, covering its several unique features.

Elements of the System

The many functions which must be performed to enable satisfactory operation depend on a number of elements—some internal to the system's circuitry and others associated with its environment. These elements may be grouped into ten general categories:

1. A very sensitive distant talking microphone with associated high-gain amplifiers.
2. One or more loud-speakers, also associated with a high-gain amplifier.
3. A line-connecting circuit.
4. Variolossers (devices in which loss can be controlled by voltage or current) with voice-operated automatic switching control.
5. Transmitting and receiving gain control circuits.
6. An automatic control for switching sensitivity.
7. Speech-continuity control circuits.
8. Talker-recognition circuits.
9. Guard circuits.
10. Reverberation control devices.

A block schematic of the major circuit elements and their manner of interconnection is illustrated in Fig. 1, and Fig. 2 shows a group of equipment assemblies used for this system.

System Stability

The microphone, centrally located within the conference room, is effectively the monaural ear of the person, or persons, not present in the conference room, and the loud-speaker (or speakers) represents his lips. For his satisfactory vocal and aural participation with all conferees within the room and regardless of their location therein, the accoustical coverage fields of these elements must completely overlap.

For the same reason, substantial electric power gains are necessary in the amplifiers of both transmitting and receiving branches of the system to insure adequate loudness at both ends of the telephone connection. Furthermore, in order to permit 2-way conversation over a normal 2-wire subscriber loop connection to a telephone central office, a hybrid coil, with an associated balancing network, is required to interconnect both the microphone and loud-speaker branches of the system with the loop. Under these circumstances, special precautions are necessary to maintain system stability against self-excitation with resultant oscillation or singing.

An idle loop singing supressor circuit, at the junction of the transmitting branch with the hybrid coil, prevents singing when no telephone connection is established to the system. This circuit con-

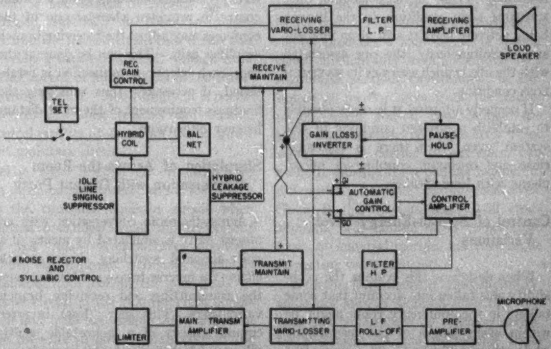


Fig. 1. Block diagram of major circuit elements for Large Conference Room Speakerphone System

---

Paper 60-1250, recommended by the AIEE Communication Wire Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960, and re-presented for discussion only at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 1, 1960; made available for printing December 28, 1960.

Luther R. Huggler is with the Bell Telephone Company of Pennsylvania, Pittsburgh, Pa.

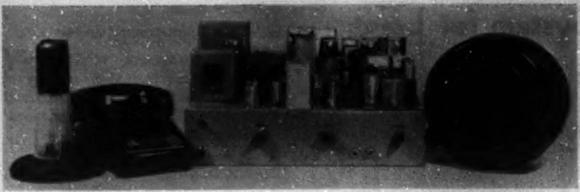


Fig. 2. Equipment assemblies for Large Conference Room Speakerphone System

sists of a network of semiconductor diodes which effectively establish an OPEN-CIRCUIT condition at this junction until they are biased for conduction by a direct current from a telephone central office or PBX (Private Branch Exchange) switchboard.

At the time a telephone connection is established, a portion of the received direct current is diverted to a set of varistors in the balancing network. The network impedance is thereby electonically adjusted to balance more closely the impedance of the telephone loop. The improvement obtained by this method, while desirable under any circumstance, will not in itself completely inhibit the establishment of singing conditions.

Complete stability at all times while the system is in use is achieved by adjusting the system internal losses so as to insure that their total, including both electrical and acoustical paths, always will exceed the total gains. To accomplish this result without effectively canceling the high gain needed in the system, complementary variable-loss devices (variolossers), having an adjustable range of about 30 to 40 db (decibels), are associated with the transmitting and receiving branches. These devices assume their MINIMUM LOSS condition in the branch being actively used at any given moment and, simultaneously, the one associated with the idle branch assumes its MAXIMUM LOSS condition.

If properly adjusted, it is rarely possible to establish a SINGING condition in this system, even when there is extremely close and optimum coupling of microphone with loud-speaker.

Control of Speech-Energy-Level Variations

When designing this system, the probability was taken into account that some talkers in the conference room might be within arm's length of the microphone while others might be 15, 20, or more feet away. Furthermore, a loud talker might be close to the microphone and a weak one at the most remote point in the room. The gain of the transmitting branch amplifiers must, accordingly, be sufficient to provide an adequate loudness in the absent party's receiver from the most remote and most soft-spoken talker in the conference room, over any reasonably high loss telephone connection. On the other hand, the output level of the transmitting branch from a loud talker close to the microphone must not be so great as to cause discomfort to the distant listener or to cause crosstalk into the telephone system plant. The conflict of these requirements is minimized by means of a thermistor limiter at the output of the main amplifier in the transmitting branch.

The acoustic output level of the loudspeaker (or speakers) in the conference room must be sufficient, of course, to be heard easily by a conferee—even by one with perhaps some slight hearing impairment and even though he might be seated at a point most remote from the loudspeaker. The absent party, however, may be a loud, average, or weak talker and may be located at the far end of low-, average-, or high-loss telephone connections. Accordingly, to minimize possibly great variations of loud-speaker output level within the conference room, a volume control is provided whereby one of the conferees may adjust the receiving branch amplifier gain. This can be done at the time each telephone connection is established, if necessary, thus satisfying the loudness requirement of the most distant listener in the room.

Simulation of Across-the-Room Conversation with Distant Party

Across-the-room conversation with an absent party is simulated by means of a voice-actuated switching device which effects the inverse transfer of loss between the transmitting and receiving branch variolossers. This gain (loss) inverter circuit consists of a monostable multivibrator which varies d-c biasing voltages applied to semiconductor diodes in the variolossers. In the AT REST condition of the system, the receiving branch variolosser is normally biased for minimum loss.

Present in the transmitting branch is signal energy from the microphone; this is sampled, amplified in a control amplifier, rectified, and applied as a positive d-c biasing voltage to actuate the multivibrator. The latter, in turn, effects reduction of loss in the transmitting branch and increases loss in the receiving branch. The initiation speeds of the switch and loss transfer are such that the distant party is rarely aware of any clipped initial syllables in the transmitted speech.

Signal energy of the distant party's speech is prevented from adversely transferring loss to the receiving branch because of the accoustic coupling between loudspeaker and microphone by including a 3-kc (kilocycle-per-second) low-pass filter in the receiving branch of the system, while a 4-kc high-pass filter is used at the input to the control amplifier. These filters enable the switching system to recognize whether speech energy reaching the microphone originates from a conferee within the room or from the distant party, and thus cause switching of loss, or its avoidance, whichever is appropriate at the moment.

Operation of the switch can be maintained at the TRANSMITTING condition by a sample of outgoing signal being taken from the main transmitting amplifier, amplified further, and then rectified and applied as a positive d-c biasing voltage to increase the control amplifier gain. This transmit maintain circuit thus assures decisive operation of the switching system by transmitted speech, even with extreme variations of signal energy from a weak and/or remote talker in the conference room.

Switching DROP BACK is avoided when no speech energy is present in the transmitting branch because of pauses between words and sentences, by using a pause-hold circuit, placed between the control amplifier and the gain (loss) inverter. This circuit momentarily maintains the switching bias on the multivibrator, awaiting a subsequent word or sentence from the talker in the conference room.

Up to this point, this speakerphone system has been in a condition to favor received speech from a distant party only when AT REST. However, when a person in the conference room initiates switching action, the distant party is, in effect, locked out. Also, when speech in the conference room has ceased, the switch may be maintained against the distant party by high-frequency (above 4-kc) ambient noise in the conference room.

Provision for BREAK-IN capability by

the distant party in simulation of across-the-room conversation takes the form of a receive maintain circuit included in the system. This circuit samples signal energy received from the telephone loop, and amplifies and rectifies it to develop a negative d-c control voltage. The sensitivity of this circuit is adjusted concurrently with operation of the loudspeaker volume control by one of the conferees. The negative voltage from the receive maintain circuit is applied at a common point in the system to oppose the positive voltage from the transmit maintain circuit. Hence, unless strongly opposed by transmitted speech, the receive maintain voltage will disable the control amplifier and cause a rapid discharge of the pause-hold circuit voltage. This action takes place with such rapidity that persons present in the conference room are rarely aware of any clipping of initial syllables in the received speech.

Addition of the receive maintain circuit, however, introduced a new operational problem. Transmitted speech signals from the conference room, which leak across the hybrid coil because of imperfect impedance matching between the loop and the balancing network, will tend to activate the receive maintain circuit in the same manner as for speech received from the distant party. Another amplifier-rectifier circuit, the hybrid leakage suppressor, therefore, must be added. Thus, if speech signals are present in the transmitting branch, these are sampled to generate a direct voltage, which disables the receive maintain circuit unless it is already strongly opposed by the voltage developed from speech signals of the distant party.

Both TRANSMIT LOCK OUT and RECEIVE LOCK OUT are deliberately designed into the system. Accordingly, an enforced politeness is necessary to allow pauses in the other person's speech before attempting to break in. These features have been necessary to assure rapid and positive switching for talkers at considerable distances from the microphone in the conference room, and also to attain as nearly as possible the simulation of across-the-room conversation with the absent party.

Ambient Noise in the Conference Room

As was intimated, the switching control elements of this system cannot distinguish between the high-frequency components (above 4 kc) in the speech of a conferee within the room and the high-frequency components (also above 4 kc) of the ambient noise, constantly or intermittently present therein. Hence, initiation or maintenance of a switch of the system to a TRANSMIT condition, resulting from such high-frequency ambient noise, might interrupt or lock out the distant party. To alleviate this situation, an automatic gain control circuit (AGC) is provided to adjust the switching system sensitivity in the presence of ambient noise.

A second rectifier at the output of the control amplifier develops a negative d-c control voltage from the high-frequency signals present in the transmitting branch. This voltage is fed to a gain decreaser (GD) section of the AGC circuit and thence back to the control amplifier input to decrease its gain sensitivity.

Concurrently, a positive d-c control voltage is obtained from the gain (loss) inverter circuit. This voltage is fed to the gain increaser (GI) section of the AGC circuit and thence also to the control amplifier input to increase its gain sensitivity.

The AGC voltages, GD and GI, normally act together to adjust the control amplifier sensitivity and maintain the gain (loss) inverter circuit in its monostable RECEIVING condition while making it immediately sensitive for switching to the TRANSMITTING condition.

To keep speech signals at either end of the telephone connection from operating the AGC circuit, opposing d-c control voltages are obtained from the transmit maintain and receive maintain circuits to suspend AGC action when speech signals are present in the system.

Experience indicates that automatic adjustment of switching sensitivity by AGC action in the presence of ambient high-frequency noise greatly improves operation of the system. But this, in itself, is not enough to prevent false or sluggish switching in some instances, because the transmit maintain circuit can develop a suspending AGC voltage from either speech or noise signals in the transmitting branch. Accordingly, it is necessary to add noise rejector and syllabic control circuits at the input to the transmit maintain circuit.

The noise rejector circuit establishes an open-circuit threshold at the input to the transmit maintain circuit when normal room noise signals only are present in the main transmitting amplifier. Thus, the transmit maintain circuit is inhibited from freezing the gain (loss) inverter circuit in the TRANSMITTING condition and from suspending AGC action in the presence of noise signals only in the transmitting branch of the system.

To permit normal operation of the transmit maintain circuit when speech signals are also present in the transmitting branch, the syllabic control circuit is associated with the noise rejector circuit. The latter detects the syllabic variation of the transmitted speech when present with the noise signals and acts to reduce or eliminate the noise rejection threshold, thereby enabling full transmit maintain circuit action

Reverberation in the Conference Room

Reverberation of speech originating in a conference room is detected by a microphone simultaneously with the speech signal received directly from the speaker's lips, and the resulting mixture may cause the signal at the distant end of the telephone connection to be completely unintelligible. Such distortion is greatly reduced if the talker remains close to the microphone, thereby maintaining a low ratio of reverberation, relative to the directly transmitted speech signal. This restriction, however, is not compatible with the conditions established for this speakerphone system.

Intelligibility, we discovered, generally suffers more when reverberation results from low-frequency than from high-frequency components of speech. Hence, a conference-room speaker can move quite far from the microphone if the gain-frequency response characteristic of the system's transmitting branch is made less efficient in transmitting low frequencies than in transmitting high ones. Accordingly, a low-frequency de-emphasis device (designated in Fig. 1 as LF Roll-Off) is provided. This device produces about 3 db of loss at frequencies near 1,000 cps (cycles per second) and about 6 db of loss per octave at lower frequencies. The resulting low-frequency roll-off produces a high degree of speech signal intelligibility from a conferee, in a reasonable acoustic environment, at distances 20 feet or more from the microphone. Such roll-off can change the tonal characteristic of a speaker's voice signal and, if carried to extremes, can make it difficult for a distant listener to differentiate between various conferees. Since, in this system, the roll-off characteristic was chosen on a compromise basis, no material sacrifice has been made respecting voice recognition.

A further improvement of intelligibility, while not always required with this system, can be obtained by ordinary acoustic treatment of the conference room walls, ceiling, and floor. Such treatment has little effect at low frequencies, but the resulting reduction of high-frequency reverberation does enhance intelligibility of transmitted speech signals.

Using a single centrally located microphone, as specified for this system, rather than several microphones dispersed within the room, also helps to minimize adverse effects of reverberation.

Conclusions

The Large Conference Room Speakerphone System gives every indication of being a highly satisfactory, commercially applicable, group-communication device. After proper adjustment to conform with the environment where it is installed and used, its operation from a customer standpoint is as simple and reliable as that of any other speakerphone system. Its operation is also compatible for interconnection with like systems and with other presently available types of speakerphone systems which may be installed at other locations.

Because of its built-in enforcement of politeness in conversational interchange, some people who have a habitual urge to break in during conversations may encounter a degree of frustration during their initial period of using the system. Experience has shown, however, that they usually adapt to this situation quickly and are highly satisfied with the entire system.

◆

Discussion

Robert A. Rivenes (Illinois Bell Telephone Company, Chicago, Ill.): While originally designed to overcome severe acoustic problems, the basic concepts of this design should not be overlooked where they apply to other problems—for example, in connecting telephone lines to radio circuits and in particular to simplex radio channels wherein the radio transmitter and receiver operate on the same frequency. The switched-loss concept could be employed for more positive switching in place of the voice-operated relay commonly used. The switching could be controlled only from the telephone-line side, since the receiver cannot distinguish local transmitter from a distant one.

While the author did not indicate what adjustments are necessary during installation, it can be assumed that microphone preamplifier gain and loud-speaker volume are adjustable. Internal adjustments also may be required to establish proper switching thresholds, once room acoustics and user requirements are set up.

Luther R. Huggler: Operation of a switching system based upon detection of frequency spectrum differences between near- and far-end talker speech signals, such as is provided in the Large Conference Room Speakerphone System, can be made quite positive and inherently more reliable than a voice-operated device, which is dependent only on the differences of speech-signal-energy levels which may (or may not) exist along a voice transmission system. The circuitry necessary to assure the existence of adequate speech-signal frequency spectrum differences at the switching point will probably involve some economic penalty, which must be balanced against the degree of improved reliability thus obtained.

The speakerphone system under discussion is equipped with internal adjustment features to adapt it for a wide range of acoustic environments within which it may be operated. These are provided in addition to the usual amplifier volume controls necessary to meet the users' loudness level requirements. If the switching principle used in this system is adapted for other uses, such as for a simplex radio channel, it is quite probable that some of these internal adjustment features could be omitted and perhaps some of the guard circuits could also be simplified or omitted.

The Subscriber Loop Bridge-Lifting Problem and Some Methods of Solution

LESTER HOCHGRAF
ASSOCIATE MEMBER AIEE

TO THE NONTELEPHONE engineer, the title of this paper may suggest difficulties involved either in raising a drawbridge or removing some dental hardware. Actually, bridge lifting is a telephone problem in exchange switching and subscriber loop layout.

Uses of Bridge Lifters

Some examples of the reasons for bridge lifting, the methods proposed for accomplishing it, and the many requirements imposed on the devices may be informative. This paper will first illustrate some applications for subscriber bridge lifting, and then describe several suitable devices, including the saturating inductor which appears to have many desirable characteristics for bridge-lifting circuits.

OFFICE-LOCATED BRIDGE LIFTERS

A doctor has a telephone in his office and wishes an extension in his home some distance away. The two telephone loops to his home and office therefore are connected in parallel, or bridged, at the telephone office; thus, calls may be originated or received at either location. Either unused loop, bridged across the one in use, can cause transmission loss and be a source of distortion to the voice currents. A bridge lifter, in this case, can be a relay circuit at the telephone office which disconnects the loop not in use, but reconnects both loops when idle to the office terminals, permitting ringing signals to reach the two lines. This type of application is shown in Fig. 1(A).

Another doctor may subscribe to an answering service that may or may not be located in the same central office area. The answering-service line will be bridged on his line and therefore will cause a transmission loss. The doctor's bridged line likewise causes a loss when the answering service takes a call.

The foregoing examples—an off-premise extension bridge and an answering-service bridge—point out the need for two different kinds of circuits intended to improve transmission by disconnecting the idle loop. The answering-service line should be completely cut off to insure privacy when the customer uses his telephone. However, in the case of the off-premise extension, such exclusion might be quite undesirable. Should the office delay in answering, the call could be taken at home. But the office would then be cut off and would be unable to

---

Paper 61-164, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 20, 1960; made available for printing November 28, 1960.

LESTER HOCHGRAF is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

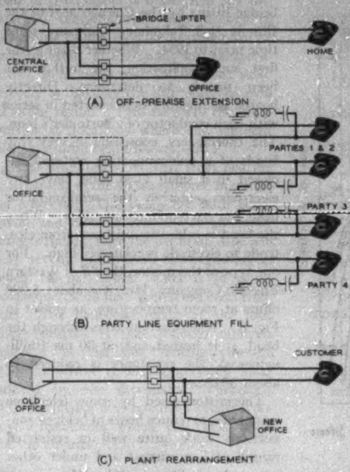


Fig. 1 (left). Applications of bridge lifters

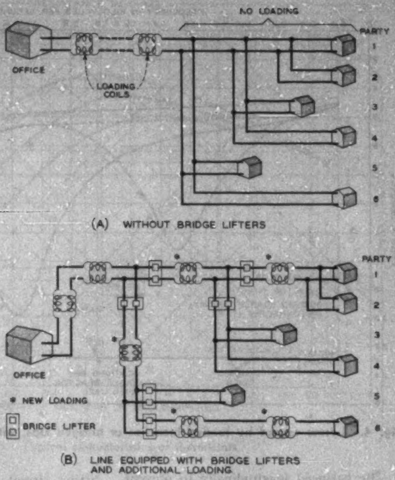

NO LOADING
PARTY
OFFICE
LOADING COILS
1
2
3
4
5
6
(A) WITHOUT BRIDGE LIFTERS
PARTY
OFFICE
1
2
3
4
5
6
* NEW LOADING
BRIDGE LIFTER
(B) LINE EQUIPPED WITH BRIDGE LIFTERS AND ADDITIONAL LOADING

Fig. 2 (right). Use of bridge lifters and increased use of coil loading on rural party lines

break into the conversation or to initiate an outgoing call. For such applications, a nonexclusion arrangement is necessary. With this latter type of bridge-lifting circuit, either location may answer a call, break in, or listen in at any time.

Bridge lifters are used also when several party-line subscribers, who are actually served by different lines, must be connected to one equipment terminal to increase the fill of central office equipment. A bridge lifter circuit, in this case, would permit bridging these lines, without rearrangement of subscribers, as shown in Fig. 1(B). For such applications, an exclusion-type bridge lifter is not permissible as it would let one subscriber use the line and exclude all others.

REMOTELY LOCATED BRIDGE LIFTERS

The possible uses for remotely located bridge lifters are greater than for those at telephone offices. For example, office cutovers or rearrangements may be effected when a group of customers, for many months connected by branch cable pairs to both old and new offices, will use the old office with a branch line to the new one bridged on. After cutover, they will be served by the new office while the old branch cable remains connected and adds loss until it can be conveniently removed. Bridge lifters at the branch point in the cable can prevent added transmission loss which, otherwise, would result from the idle branches. Such an arrangement is shown in Fig. 1(C).

Another important application for bridge lifters, shown in Fig. 2(A), is multiparty rural lines, where transmission losses sometimes exceed the objective because of the extra-long lines and the added loss caused by many subscriber branches. In the line section where branches to various parties are found, coil-loading the pairs would cause severe frequency distortion, emanating from reflections from the idle or open end of a loaded branch. However, if bridge lifters are inserted at each branching point in both main line and branch as shown in Fig. 2(B), loading can be carried nearly to the end of the branches without causing distortion and with a worth-while reduction in transmission loss. Thus, a double benefit is obtained: an increased amount of loading, coupled with a reduced branch-line bridging loss.

Benefits of Bridge Lifting

The effective losses caused by idle branch circuits are significant, as shown by Fig. 3. A good rule of thumb for nonloaded cable branches on a nonloaded pair is 0.25 db (decibel) per kilofoot of branch. Each 6,000 feet of bridged tap thus adds a 1.5-db effective loss, which is greater at high-voice frequencies than at low. A nonloaded branch on a loaded line adds even more loss and lowers the cutoff frequency of the line. Moreover, when the idle branch is loaded, the loss and distortion can be quite severe, as is shown in Fig. 4. This is because of the large phase shift of the reflected wave from the open end of the loaded branch. Hence, a successful bridge lifter reduces restrictions on the length of bridged taps, and improves transmission up to 3 or 4 db in cases where provision of good transmission is otherwise difficult.

OPERATIONAL REQUIREMENTS

A bridge lifter to be used in present telephone plants should:

1. Pass and withstand 20-cycle ringing currents.
2. Switch to low transmission loss for direct current values created by an off-hook condition, but present a high impedance on-hook with leakage currents found on normally well-maintained lines.
3. Permit dial pulsing without serious distortion.
4. Have low d-c resistance or small voltage drop when a branch is in use.

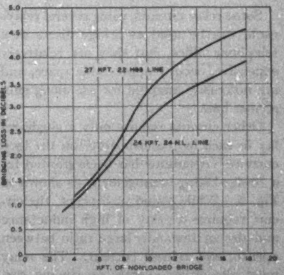


Fig. 3. Effective losses caused by nonloaded branches

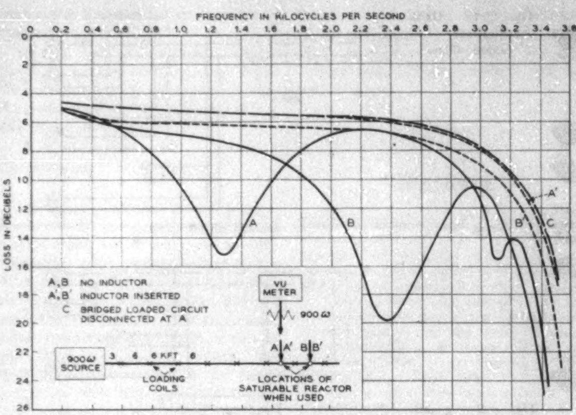


Fig. 4. Losses caused by loaded branches or bridged taps, with and without bridge lifters

5. Be balanced longitudinally to prevent generation of power-induced noise.

6. Not generate harmonics with 60 cps (cycles per second) metallic voltage across the pair.

7. Not require a source of direct current other than that supplied by the line pairs.

8. Have low cross-talk when mounted in close proximity to other bridge lifters.

9. Withstand lightning and switching surges.

10. Operate satisfactorily in an outdoor environment.

If the electrical characteristics of the device permitted several bridge-lifting units to be placed in tandem without transmission degradation, they would be useful on multiparty lines. However, quite tight electrical characteristics would be required.

Relay Circuits as Bridge Lifters

Some relay circuits, using several relays each, can be used as bridge lifters at a telephone office, but they are relatively expensive, occupy appreciable space, and are not easily adapted for use along subscriber lines.

Other circuits require only a single relay in each loop and operate on the loop current. A patent[1] has been granted for a bridge-lifting arrangement as shown in Fig. 5. While sound in concept, the circuit requires a relay of high inductance and also an unusually large ratio between operate and release current. The relay must also either operate on ringing currents or else the ringing voltage will have to be applied on the line side of the bridge-lifting circuit. Thus, if ringing cannot pass through the relay, it would restrict this type of circuit to office locations.

Since successful relay circuits are not small or low in cost, a continuing search has been made by telephone engineers for nonlinear devices which can perform bridge-lifting functions. A few such devices come very close to meeting the many and conflicting requirements imposed, and can be used with restricted applications where some of the conditions are not controlling.

Thermistors as Bridge Lifters

The use of thermistors, connected in series with the line, was proposed for bridge lifting by the Chicago area of the Illinois Bell Telephone Company some time prior to 1954. Whether this was the first such application, the author has been unable to determine. Directly heated thermistors were inserted in series with each conductor of a customer's loop. The thermistors, consisting of very tiny beads of a semiconducting material enclosed in a small glass tube, have two electrodes entering the semiconductor bead, but not touching each other. When the bead is cold, the resistance from electrode to electrode is relatively high. For one type of thermistor, the *8B* of Western Electric Company, Inc., it is about 5,000 ohms at room temperature, as shown in Fig. 6. When current flows through the bead, it is heated and, at 30 ma (milliamperes), the resistance is reduced to about 50 ohms.

Thermistors used by some telephone companies to reduce losses of bridged connections work quite well in restricted ranges of line current and under other favorable conditions, such as low power induction. A pair may insert 100 ohms in a branch when in use, and 10,000 ohms when not in use.

Temperature Effects

Since the thermistor is a temperature-dependent device, the effects of ambient temperature, especially in the cold condition and warm-up time, are important.

The cold (no current) resistance of an *8B* thermistor is reduced to about 2,500 ohms at 120 F (degrees Fahrenheit), and is raised at 32 F to 15,000 ohms. At an ambient of 120 F, the bridging loss of a

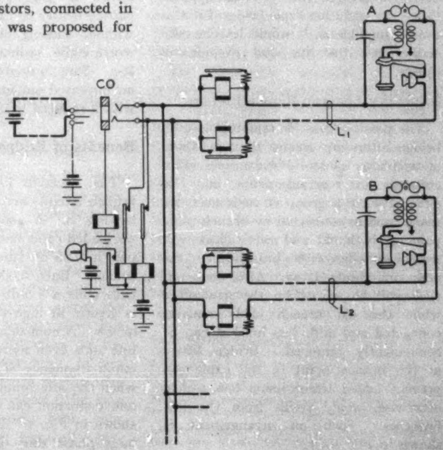


Fig. 5. Relay bridge lifter, Patent no. 2,039,413

pair of thermistors is more than 1/2 db; at low temperatures, time delays in switching caused by slow heating may be encountered; at extremely low temperatures, some units may fail to heat sufficiently to "go over the hill" into the conducting region. This restricts the thermistor to locations where the ambient does not drop below about −20 F.

The time delay resulting from heating and cooling in the *8B* is measured in tenths of a second. This delay has both beneficial and adverse effects. It adversely affects ringing since the warm-up time will clip several pulses from the beginning of each ringing spurt. This is of some consequence on lines which use code ringing. The beneficial effect of the long time-constant is that induced 60-cycle current in the line flowing through the thermistor is not appreciably modulated by the thermistor. Therefore, the thermistor does not itself produce 60-cycle harmonics which can cause noise in the circuits.

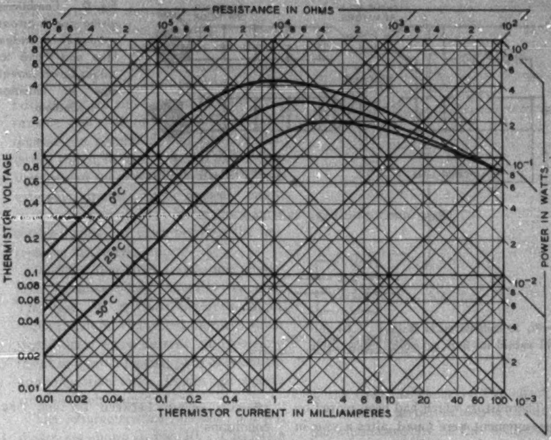


Fig. 6. Typical characteristics of thermistors

EFFECT ON DIAL PULSING

The effect of thermistors on dial pulsing is to reduce the range of a subscriber loop by 450 ohms, which is a substantial penalty. The reason for this reduction lies not so much in the time delay of the thermistor as in the fact that the average current when pulsing is less than the steady off-hook current, and the thermistor resistance thereby increases.

LONGITUDINAL BALANCE PROBLEM

To prevent 60-cycle induced power voltages and harmonics from causing noise in the metallic circuit over which speech currents flow, nearly equal resistance must be inserted in the tip and ring conductors. These resistances must be held to a few ohms difference in the talking condition with line current flowing in the thermistors. It is also essential to have less than 5% difference in resistance in the low-current or idle-circuit condition since the idle branch is bridged on another branch, which may be in use, and the idle thermistors can thus introduce noise to the branch being used.

Thermistors, as manufactured, do not possess this degree of uniformity in resistance from unit to unit. It has been found necessary to match or pair thermistors according to resistance to meet longitudinal balance requirements, discarding a fairly high proportion which could not be paired within the rather narrow limits.

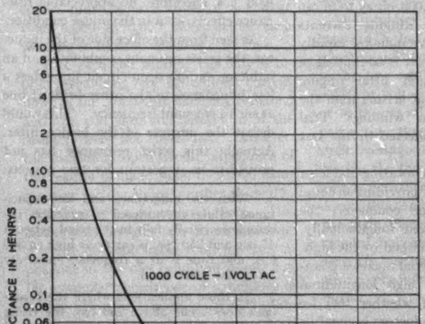
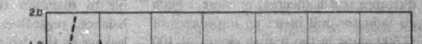
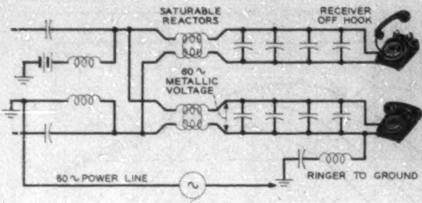


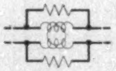
Fig. 7. Inductance of 1574A inductor as a function of direct current in windings

Fig. 8 (below). Bridging and insertion losses of 1574A inductors in telephone lines





NOISE PRODUCTION ON IDLE PAIR



SATURABLE INDUCTOR WITH HIGH-RESISTANCE SHUNTS

Fig. 9. Conditions giving rise to power voltage across bridge lifter and shunts to reduce power current flow in inductor windings

Thermistors which had been paired by measurement were found, after a year in service, to have drifted apart in resistance, and an appreciable per cent of pairs were out of balance. Some of these thermistors were also quite high in resistance in the cold condition. The change was attributed to the effects of line current and especially to ringing current on the thermistor bead. Laboratory tests did indeed show damage to some thermistor beads caused by ringing currents. However, other units which had been put on the shelf for one year also drifted out of balance.

Because of this propensity to change characteristics with time and because of the marginal ON to OFF resistance characteristics, the search was continued for a more generally useful device.

Saturating Inductors for Bridge Lifting[2]

The insertion of a pair of inductors or a balanced-winding inductor in the tip and ring conductors of a branch will effectively reduce the bridging loss if the inductance is 3 h (henrys) or more. Such an inductor will insert an impedance of $j7{,}500$ ohms in the branch circuit at 400 cps and a higher impedance at higher frequencies. Of course, when the branch is put into use, this series inductance must be switched out. This action can be accomplished by a self-saturating inductor such as the Western Electric *1574A*, which was developed for this purpose. The inductor, wound on a permalloy tape core without an air gap, provides 18 h with no direct current, 6 h with 1 ma direct current, and only 0.025 h at 20 ma. This large variation in inductance is illustrated in Fig. 7, and provides a very effective switch between ON and OFF conditions.

Since the d-c resistance of both tip and ring windings is less than 25 ohms, the reduction in d-c flow to the telephone set is small. The insertion loss of the device for d-c line currents 10 ma or greater, is shown in Fig. 8, as is the bridging loss of a branch circuit equipped with a *1574A* inductor for leakage current of less than a few ma. These values are quite satisfactory for existing telephone sets on reasonably well-maintained lines where leakage is held to 50,000 ohms from tip to ring conductors. Ringing currents saturate the core of the inductor rapidly on each pulse; thus, there is no clipping of the ringing spurt and little ringing penalty other than that arising from the d-c resistance of the windings. Dial pulsing is also little affected by the fast action of this device.

What makes the saturating inductor so attractive is that, by providing an equal winding for each line conductor, the circuit can be balanced longitudinally. The tight coupling furnished by the high-permeability core material makes it possible to maintain this high longitudinal balance, regardless of whether the d-c leakage in the two conductors is equal or not. Longitudinal alternating currents also do not magnetize the core and are, therefore, not modulated by the device.

An inductor inserted in each conductor of a pair is subject, on some lines, to induced power currents which flow in these conductors through the office battery and ground paths and return to a remote ground. With a balanced winding arrangement, the equal currents in each conductor do not magnetize the core and thus cannot generate harmonics. The core is magnetized only by the difference or loop current, which is a function of both the induced voltage and the imbalance caused by office equipment or by telephone set connections. However, devices inserted separately in each conductor are subject to the full longitudinal current flow and, therefore, are likely to cause modulation noise.

Finally, maintenance of saturating inductors is minimal. They are essentially of the same construction as loading coils which have served in cable pairs for years, withstanding switching surges, ringing currents, and lightning surges, with a remarkably low trouble rate.

The saturating inductor, therefore, appears as a favored device for bridge lifting. However, cognizance should be taken of the possible modulation of induced 60-cycle metallic circuit current and of resonance between the inductor and the idle-line capacitance, as discussed in the succeeding sections.

LINE-RESONANCE EFFECTS

Since an inductor stores energy and does not dissipate it, there can be marked resonance effects between its inductance and the capacitance of an idle- or open-circuit loop. On the other hand, when the loop is terminated by an in-use telephone set, a dissipative termination is supplied to the line and the inductance of the saturating inductor is driven below 0.03 h. Therefore, no appreciable resonance effects occur in the in-use condition.

As seen from the office side of the inductor, the series-resonant combination of an inductor and an open-circuit line offers a low impedance to the other bridged line at such a resonant frequency. This would defeat the purpose of the bridge lifter. Actually this series resonance has not proved to be important for two reasons:

1. With the inductance and line-capacitance values encountered in practice, the resonance usually falls in the band between 45 cps and 250 cps, is rarely as high as 300 cps, and thus is at a frequency generally below the voice band.
2. In those applications where the resonance does occur above 250 cps, the rising resistance of the inductor, largely induced by eddy currents, is effective in reducing the loss. Thus, a typical *1574A* inductor has an a-c resistance of 1,000 ohms at 400 cps in the low-current condition, and the minimum bridging loss in the narrow resonant band is less than 4 db. The effective resistance can be further raised by use of thicker core tape, but this is undesirable since it decreases the change in inductance from low- to high-current conditions.

When the resonant frequency between the inductor and open-circuit branch is close to 60 cycles, this can increase the metallic power voltage across the inductor above the open-circuit value and, if sufficiently high, drive enough current through the inductor to cause modulation, and therefore noise, in the connected branch. The situation is diagrammed in Fig. 9.

Reducing Power Current in Windings

The power-frequency voltage across the inductor can be reduced by avoiding ringers connected directly from tip or ring conductors to ground. With ringer connections to ground, the open-circuit power-frequency voltage across the inductor may be one-half the voltage from each conductor to ground. With ringers connected between tip and ring conductors and not to ground, or with cold-cathode gas tubes inserted in a grounded ringer connection, this ratio can be reduced to one-tenth or less. Since these remedies, although effective, can be expensive to install on existing lines, other cures are necessary if this application is to gain universal appeal.

Some of the induced power current from the reactor windings can be diverted along shunt paths connected across each winding. Resistance shunts are especially effective in the near-resonant situations where noise generation is the highest. With a *1574A*, these shunting resistors must have a value of not more than 4,000 to 6,000 ohms across each winding to be fully effective.

Since each coil winding has only a 10-ohm resistance, very little direct current flows in the 5,000-ohm shunting resistors, and the coil saturates almost as if the resistors were not connected. However, the maximum open-circuit impedance of the reactor is made 8,000 to 10,000 ohms resistive by the shunts, rather than $j15{,}000$ ohms or more without them. This causes more bridging loss than is desirable. A value of 8,000 ohms total for the impedance in the OFF condition would not, for example, permit use of a number of bridge lifters in tandem on a party-subscriber line. The inductor design, however, can be modified quite easily in order to raise the inductance. This will permit higher-value bridging resistors and thereby reduce the bridging loss of the treated branches.

Solid-State Devices as Bridge Lifters

While the thermistor is a solid-state device comprised of a semiconducting material, the term "solid-state device" is usually reserved for junction devices such as rectifier diodes and transistors. Transistors used for switching purposes have OFF and ON states under control of a third electrode, and p-n-p-n diodes can be fabricated with a threshold voltage in both backward and forward directions to provide OFF and ON states. The state depends on current flowing through the device which, in turn, depends on applied voltage and the current permitted by the external circuit. Such devices merit consideration as bridge lifters since very sizable ratios can exist between the impedance in the two states.

Diodes rather than triodes are preferable because they simplify circuits. The bias for a third electrode might require a shunt-connected bleeder, which would draw extra current, and possibly result in a cross-ringing path between tip and ring conductors. Diodes can operate without such shunts.

The characteristics of one p-n-p-n device are shown in Fig. 10. This has a threshold in the forward direction of 20 volts, with the fully ON condition not obtained until the d-c bias is 10 ma. Such devices are essentially unidirectional. Since the battery voltage may be reversed on some loops and both positive and negative ringing pulses need to be passed, it is necessary either to place two diodes in series in each conductor poled in opposite directions, or in parallel and oppositely poled.

The problem when using diodes is to insure: (1) that the longitudinal balance is not upset by unequal leakage currents in the two conductors, (2) that there is sufficient voltage across the line with two or more sets off-hook to switch into the conducting range, and (3) that the 60-cycle voltage across the diode never reaches the threshold value, causing it to switch periodically to the ON condition. Again, this can be aided by shunts around the diode units and especially by a tightly coupled balanced-inductor shunt, which serves to maintain balance and to by-pass the power-induced currents from the diodes.

Disadvantages of Bridge-Lifting Devices

Bridge lifters are not perfect switches; moreover, their introduction in the telephone plant—as is true of any new device—almost inevitably causes some modification of prevailing practices. One conflict is with open-circuit testing procedures, used either to measure line leakage or to check continuity through to the customer set by observing the charging of the ringer condenser. The effect depends on the particular testing circuit and the characteristics of the bridge-lifting device.

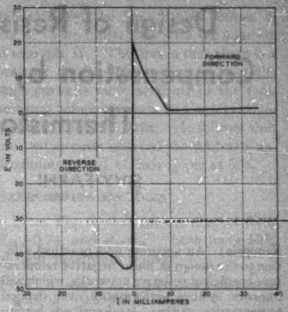


Fig. 10. Typical characteristics of p-n-p-n diode which could be used for bridge lifting

Another conflict occurs on party lines equipped with telephone sets with switch-hook plungers, which permit listening in without drawing line-current. This type of set is occasionally used on rural party lines so that subscribers can determine if the line is idle before dialing. The bridge lifter defeats this operation by requiring line current on a branch before the set can receive speech from an established connection.

Bridge lifters operate best when the line leakage and induced current are low. Thus, their use favors good maintenance of subscriber lines to reduce both leakage and equipment imbalances, the latter being caused primarily by grounded ringers on lines exposed to power induction.

Not many more requirements can be heaped on the already burdened device. To meet the demand for a bridge lifter, which will operate with telephone sets drawing currents as low as 10 ma, will be difficult, but the last straw would come if any proposal for tone-ringing is adopted using signals at speech levels. Obviously, the aim is to develop devices with universal application to telephone circuits so that bridge lifters can be used on a routine basis without consideration of induced power voltages, length of branches, or type of telephone set.

References

1. Telephone Line, Oliver Henderson. *U. S. Patent no. 2,039,413*, May 5, 1936.

2. Reduction of Transmission Loss in Bridged Subscriber Loops, Lester Hochgraf. *U. S. Patent no. 2,924,667*, Feb. 9, 1960.

Design of Resistive Temperature Compensation by Single and Multiple Thermistor Networks

SIVY FARHI
ASSOCIATE MEMBER AIEE

STANLEY GROVES
NONMEMBER AIEE

Synopsis: An accurate and practical solution to the design of linear negative temperature coefficient resistive networks, which will improve system performance over defined temperature ranges, is described herein. The basic circuit is a nonlinear negative temperature coefficient semiconductor, generically known as a thermistor, in parallel with a shunt. The temperature range of utilization of the basic circuit is about 60 C (degrees centigrade). Two or more of these basic circuits are connected in series to provide nearly linear characteristics over an extended temperature range. Graphical presentations of the basic circuit characteristics are used in an original manner, the graphs being easier to use and more complete than those previously published. Proper selection by a graphical method of the basic circuits connected in series, for wider temperature range use, is explained, and an analytic discussion tells how the graphs are constructed. The errors caused by circuit tolerances and self-heating, and the selection of thermistor materials, sizes, and shapes are outlined. Typical examples illustrate the use of the graphs and techniques. Application of the networks to temperature compensation and temperature measurement, as well as extension of the techniques to nonlinear applications, receive careful thought.

THE BASIC circuit of a thermistor in parallel with a shunt is in wide use and familiar to many engineers. This circuit, detailed in Appendix I, has the limitation of being most effective over a temperature range of only 60 C.

Basic Circuits in Series

By using two or more such circuits in series, each effective over different sections, this limitation is invalidated by extending the temperature range of linear negative coefficient thermistor circuits. The graphs for these circuits are plotted in such a manner as can be interpolated easily, and are constructed with a minimum of calculations. This method of graphical presentation is superior to the methods in present use.

The slope versus temperature of these circuits are also calculated and graphed. Then slope-temperature characteristics are replotted so they also may be interpolated easily. Although slope-temperature curves previously have been neglected, they are the key to proper selection of the circuits to be used in series. The graphs described simplify the matter of choosing parameters for the circuits in series so that their addition will result in a substantially constant slope over the temperature range of interest. Thus equipped with a set of characteristics for the various thermistor materials in use, accurate circuits may be prescribed in a matter of minutes. By utilizing similar techniques, nonlinear resistance-temperature characteristics can be obtained.

Graphical presentation of the basic thermistor circuit's resistance-versus-temperature characteristics as in Fig. 1 has the advantage of visual aid simplification. Disadvantages are that the characteristics cannot be interpolated readily for various values of R_o, and observers often mistakenly assume some characteristics to be much more linear than others—curves with shallow slopes being the most deceptive. Fig. 2 presents the same information, is easier to calculate and plot, and is readily interpolated. The characteristics of Fig. 1 can be plotted with the information in Fig. 2 for any value of R_o.

A cursory look at Fig. 1 is sufficient to show how the series connection of properly selected basic thermistor circuits will result in a substantially linear resistance-versus-temperature characteristic over wide temperature ranges. Various values of R_o can be selected by trial and error, but by development of slope-versus-temperature curves such as shown in Appendix II, better techniques that are simple, rapid, and accurate are evolved.

Each basic circuit can provide a nearly linear negative slope over a temperature range of about 60 C. The necessary number of circuits in series is approximately equal to the nearest integer of the temperature range in C, divided by 60. If, for example, a linear negative slope over a temperature range of 120 C is wanted, then from Fig. 3, two values of R_o are found which have about 50% of their slope at the mid-temperature point. The addition of these two circuits, when corrected for their differences in maximum slopes, will provide essentially a constant slope over the entire range. The method is best illustrated by considering an actual case.

Example I

A negative slope of 1 ohm per C over a temperature range of 0 C to 120 C is desired. Fig. 3 shows that at the mid-temperature of 60 C, the basic circuits which have an R_o value of 37 and 0.93 have 50% of their maximum slopes.

Fig. 3 shows that the maximum slope of the basic circuit $R_o=0.93$ is -0.01125 ohms per C per shunt ohm, and for

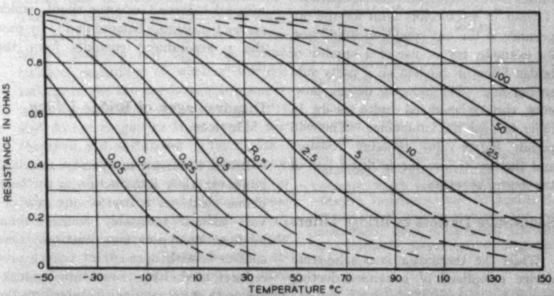


Fig. 1. Resistance versus temperature for basic pu circuit—GE (General Electric Company) Grade 1

Paper 61-189, recommended by the AIEE Indicating and Integrating Instruments Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted September 6, 1960; made available for printing December 28, 1960.

SIVY FARHI is with Hughes Aircraft Company, Culver City, Calif., and STANLEY GROVES is with Eclipse Pioneer Division, The Bendix Corporation, Teterboro, N. J.

$R_o=37$ is -0.0638 ohms per C per shunt ohm.

A relatively accurate and quick solution would be

$$R_{sa}=\frac{1}{0.01125}=88.8 \text{ ohms}$$

$$R_{ssa}=0.93R_{sa}=82.6 \text{ ohms}$$

$$R_{sb}=\frac{1}{0.0638}=156.7 \text{ ohms}$$

$$R_{ssb}=37R_{sb}=5{,}800 \text{ ohms}$$

This type of problem can usually be solved even more accurately if the residual slopes of each network are considered. From Fig. 3, it is plain that for $R_o=0.93$, the maximum slope occurs at 18 C. However, at 18 C the basic circuit, where $R_o=37$, still has a slope equivalent to 11% of its maximum slope. Similarly, the basic circuit where $R_o=37$ has its maximum slope at 120 C; at that temperature, the other basic circuit (where $R_o=0.93$) has a residual slope of 8%. This suggests that two networks should have been selected at about their 55% rather than 50% slope points.

EXAMPLE II

The basic circuits at 60 C ($R_o=1.1$ and $R_o=32$) have 55% of their maximum slopes, and this second solution is more accurate than the first.

The maximum slope of $R_{oa}=1.1$ at 22 C is -0.01.

The maximum slope of $R_{ob}=32$ at 115 C is -0.00625.

The residual slope of $R_{oa}=1.1$ at 115 C is 11%.

The residual slope of $R_{ob}=32$ at 22 C is 18%.

If the circuits were added and each circuit had the same maximum slope, the result would be 118% slope at 22 C, 110% slope at 60 C, and 111% slope at 115 C. The average slope of these three points is 113%.

Accordingly

$$R_{sa}=\frac{1}{1.13\times 0.01}=88.5 \text{ ohms}$$

$$R_{ssa}=1.1\times 88.5=97.4 \text{ ohms}$$

$$R_{sb}=\frac{1}{1.13\times 0.00625}=141.5 \text{ ohms}$$

$$R_{ssb}=141.5\times 32=4{,}520 \text{ ohms}$$

Fig. 4 graphically illustrates the slope addition of previous examples. Individual slope curves are extracted from Fig. 3. The first solution is shown by the solid lines. Addition of the slopes of the 2-series networks is relatively constant over the temperature range of 0 C to 120 C. The first solution, where the two circuits were joined at the 50% slope points, is sufficient for most applications. Although the average slope is greater than 100%, it requires less than a minute's work to determine the percentage slope at the peaks. The first solution would be more accurate if the average slope were considered 106%, and the resultant network shown to have its parameters divided by the factor 1.06.

Selecting the two circuits at the 55% slope points results in a decrease in the dip of the slope curve. The second solution, more accurate than the first, is indicated by the long dashed curve in Fig. 4. Two faults exist with the solution of example II. The first is that there is a peak at 30 C. The second is that the slope curve is more accurate in the temperature range between 70 C±60 C than it is near the 60 C value. It follows that the two circuits selected should have approximately 55% of their slopes at 50 C.

EXAMPLE III

From Fig. 3, it is shown that at 50 C $R_{oa}=0.74$, and $R_{ob}=21$; both have 55% of their maximum slopes.

Circuit *A*, where $R_{oa}=0.74$, has a maximum slope of -0.01665 at 13 C.

Circuit *B*, where $R_{ob}=21$, has a maximum slope of -0.0071 at 100 C.

Circuit *A* has a residual slope of 14% at 100 C.

Circuit *B* has a residual slope of 18% at 13 C.

Circuit *B* has a residual slope 4% higher than circuit *A*. The peak at the lower temperature can be reduced if circuit *A*'s maximum slope is 96% of circuit B's. The resulting slopes would then be 114% at 13 C, 108% at 50 C, and 112.5% at 100 C. The average slope is about 112%. Fig. 4 shows the results, as indicated by the short dashed lines. The solution is

$$R_{sa}=\frac{0.96\times 1}{0.01665\times 1.12}=51.4 \text{ ohms}$$

$$R_{ssa}=0.74\times 51.4=38.1 \text{ ohms}$$

$$R_{sb}=\frac{1}{0.0071\times 1.12}=126 \text{ ohms}$$

$$R_{ssb}=21\times 126=2{,}640 \text{ ohms}$$

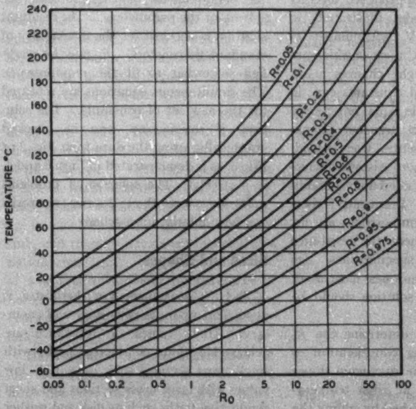


Fig. 2. Temperature versus R_o for constant values of resistance of pu circuit—GE, Grade 1

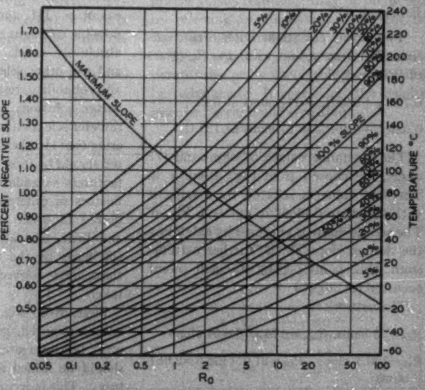


Fig. 3. Temperature versus R_o for constant values of % maximum slope and maximum slope versus R_o for pu network—GE, Grade 1

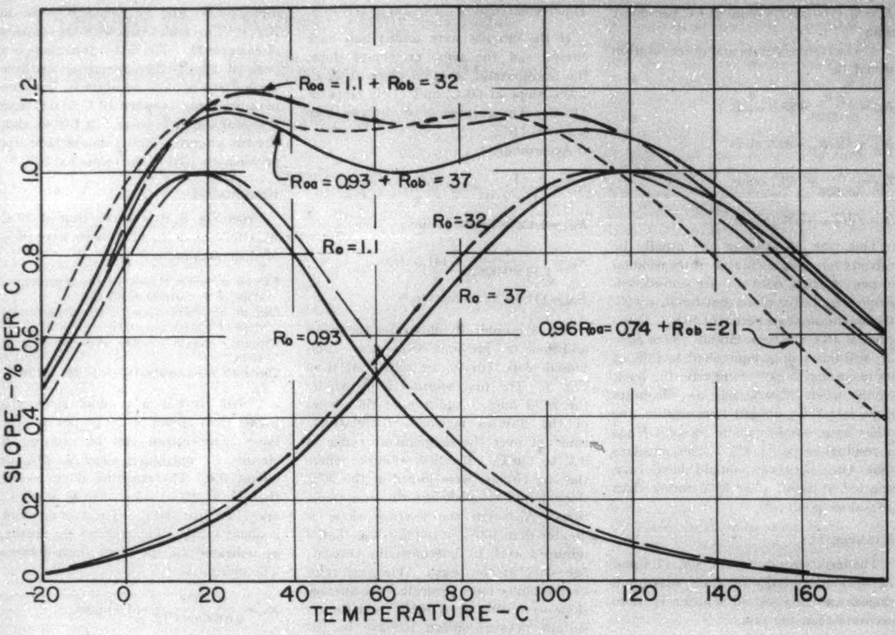


Fig. 4. Comparison of slope additions, examples I, II, and III

The resistance-versus-temperature characteristics of the three previous solutions are shown in Fig. 5. However, an improvement in results is possible, as shown in example IV.

EXAMPLE IV

Two linear equations with coefficients from Fig. 2 are set up, using the R_o values of 0.74 and 21

$0.71R_{sa}+0.99R_{sb}=177$ ohms at 0 C

$0.02R_{sa}+0.45R_{sb}=57$ ohms at 120 C

The resistance values, 57 ohms and 177 ohms, are chosen from Fig. 5, curve 3. The first value is selected by a tangent with the desired slope to curve 3, Fig. 5. The second value is obtained by extending the desired slope to the further extreme from this point. (The tangent point occurred at 120 C.)

The network is

$R_{sa}=73$ ohms, $R_{25a}=0.74\times73=54$ ohms

$R_{sb}=126.5$ ohms, $R_{25b}=21.0\times126.5=2{,}657$ ohms

The results are plotted in Fig. 6. Values of resultant characteristics are shown at 10 C intervals and are compared to the ideal straight line.

The first two solutions illustrate the technique development. In most cases, the solution can be approached directly through the methods illustrated in examples III and IV. A number of possible solutions exist, none of which are absolutely exact. Slight changes in R_o and the other selected constants can, in all probability, result in equal or possibly better solutions. This freedom of parameter selection is essential when considering that the thermistor resistances selected should be those included among manufacturers' standards. Frequently, thermistors of different compositions are required in order to meet the resistance and heat dissipation specifications. On the basis of the techniques illustrated, these steps toward a solution should be taken.

First, from Fig. 3, determine the R_o values required, giving consideration to what percentage of the maximum slopes should be added and at what temperature. This is accomplished by adding the slopes, or resistances, and bracketing the mid-temperature and the temperature extremes. When R_o values are tentatively selected, a set of simultaneous equations, using constants from Fig. 2 or Fig. 3, and the constants determined by the requirements of the problem, are solved for the parameters. The resultant solution is checked for the availability of standard thermistors. R_o may be modified somewhat to fit the requirements. The simultaneous equations are repeated for the new set of constants. The solution, if satisfactory, can be checked graphically, using the data from Fig. 2 or Fig. 3, as is demonstrated in Figs. 6 and 4, respectively. The solution of problems requiring more than 2-series networks can be similarly approached.

Sizes and Shapes

The R_{25} resistance of a thermistor is dependent upon its resistivity and geometry. High values of resistance are achieved by long conductive paths with small cross sections such as a rod, low values with large cross sections and short conductive paths such as disk and washer types. Various material resistivities are available. For example, Keystone Car-

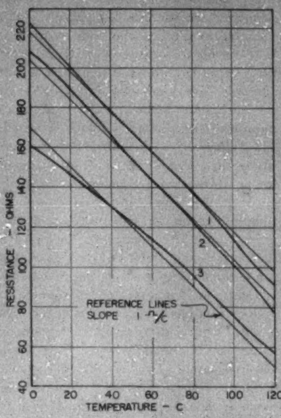


Fig. 5. Comparison of resistance-versus-temperature characteristics, examples I, II, and III

bon Company has 34 standard thermistor materials with resistivities from 196 ohm-cm (centimeters) to 97,000 ohm-cm.

Shape and mounting determine the dissipation constant of thermistors. Shapes with large mounting surfaces and short heat-conductive paths (thermistors are poor heat conductors) will have the highest dissipation constants, while washers and similar contours have dissipation constants about twice that of lead-mounted disks of about the same dimensions. For example, General Electric Company's *WA 753* and *D 751* are almost identical in size, and the dissipation constants are, respectively, 57 and 25.5 milliwatts per C. Thin washers have higher dissipation constants than thicker washers of the same diameter and material. Rods rely on the leads and surface area for cooling. This is also true for disks with leads. However, disks have short heat-conduction paths in the body of the thermistor. For this reason, most rods have lower heat-dissipation constants than disks. Washers, and some disks, are frequently mounted directly to a heat sink, and thereby have even higher dissipation constants. The heat sensitivity of a washer to ambient temperatures is better than that of rods, because of the geometry and heat-transfer characteristics.

The resultant current distribution of the thermistor is also important. A large diameter rod (or disk with leads), heating up during operation, will develop a hot inner core, which will cause the current to concentrate in this low-resistance path with a corresponding further temperature rise. This again lowers the resistance until a condition of equilibrium is reached. Such a phenomenon will occur when the heat dissipation exceeds a safe value for the conductive path, and may also take place in washers of large diameter when sections perpendicular to the current path cannot be maintained at the same temperature.

Once a hot spot develops, it frequently avalanches. The authors experienced a situation where a high-current density in a washer-type thermistor burned a neat hole through the washer. Thermistors with geometries, where the conductive paths are of different lengths, will have similar problems. The current will concentrate in the lower resistive path, severely limiting the dissipation constant. The authors recommend, on the basis of their experience, the use of the highest temperature coefficient materials which will provide the resistance required in a geometry capable of meeting the heat-dissipation requirements.

Selection of Thermistor Materials

Use of thermistors with the highest temperature coefficient will yield the lowest total series resistance. This is a highly desirable characteristic, resulting in minimum thermistor sizes, self-heating, and changes in circuit performance.

The low-temperature coefficient thermistors do have advantages, foremost of which is a more constant temperature coefficient over a given range. The second is lower resistivities. These thermistors are frequently an advantage in single networks because of their wider temperature range. Very often, thermistor materials are selected for their resistivity and not their temperature coefficient.

With the techniques described in this paper, low-temperature coefficient materials are not advantageous because of the improved method of extending temperature ranges. For this reason, the authors chose for illustration the material of General Electric's grade *1*. Thermistors with higher temperature coefficient are now readily available. For example, Keystone's 34 standard thermistor materials, previously mentioned in connection with resistivities, have temperature coefficients covering the range from −3.73% to −5.25% per C at 25 C.

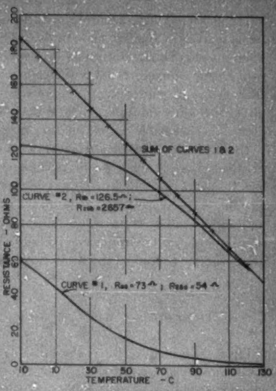


Fig. 6. Graphical illustration, example IV

Self-Heating

Assuming a constant current through the network at all temperatures (most probable case), the power dissipation in each thermistor of the estimated network may be determined. From this information, the temperature rise of each thermistor versus resistance may be predicted by referring to dissipation coefficient of the thermistor material.

This indicates the corrected temperature range for which the circuit should be redesigned. If necessary, the results may be checked and repeated.

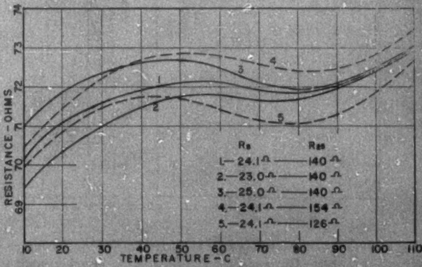


Fig. 7. Effects of parameter variation, example VI

In all probability, accurate results can be achieved in one or two trials. The dissipation constant depends on network component geometry. Manufacturers of thermistors provide dissipation constants under specific conditions. These may be used as a reasonable approximation, prior to the construction of prototype networks. Further possible changes in parameters usually will be small. However, the authors' experience with these networks indicates a small trim of the shunt resistors is all that is required. This is a quite simple procedure in most applications.

Circuit Tolerances

Fig. 7 graphically shows the effects of both shunt and thermistor variations from calculated values of the simple network case covered in Appendix I. Curve 1 is the network plot as specified, 24.1-ohm shunt and 140-ohm thermistor. Graphically illustrating the effects of parameter variation, curves 2 and 3 show the effects of ±4% change in shunt resistance, and curves 4 and 5 the effects of ±10% change in thermistor value. As would be expected, the shunt variation has its greatest effect at the lower temperature, while the thermistor variation has its greatest effect at the higher. Most important are the relatively minor change in temperature range, the general shape of the curve, and the nominal value of the entire network. The shunt tolerance is more critical than that of thermistors. Fortunately, shunts are easily obtainable with tight tolerances. The 10% variation of thermistor value does not appreciably change the curve shape, and the mid-temperature resistance of the entire network changes only about ±1%. Thermistors are usually manufactured to a 10% resistance tolerance. The need for exact value or closer thermistor tolerances should be examined most critically by designers.

The temperature coefficient of the thermistor material frequently has a tolerance on the order of ±2.7%, or about 0.018% of the beta coefficient. This does not have any appreciable effect on the circuit performance if the nominal resistance value is measured at the proper temperature. For greatest accuracy, the thermistor resistance should be specified at the temperature where the thermistor is predominately functioning. The effect of specifying the resistance value at 25 C, when the thermistor is functioning at 70 C, is similar to increasing the tolerance on the nominal 70 C resistance value an additional 20%. This error increases rapidly with a greater temperature difference.

Conclusions

The method of plotting the resistance-temperature characteristic, as shown in Fig. 2, presents a universal characteristic for the thermistor material composition. The availability of such characteristics in published form should eliminate a major part of the calculations now performed in the design of thermistor circuits. The availability of the slope characteristics, as plotted in Fig. 3, further simplifies the design of thermistor circuits, and is frequently essential to the rational design of multiple-series networks.

The network-addition advantage is demonstrated and should prove to be a valuable tool, as the methods of application are simple. Although trial and error exist, and there are an infinite number of "correct" solutions, logical application of the techniques shown in the examples minimize the number of steps required. The superiority of multiple-series networks, in overcoming the limitations of temperature range and accuracy of a single network, is evident, and ease of application and calculation will justify their frequent use. . . The relatively inexpensive thermistors and shunt resistors available can become useful low-cost improvements in system and component performance.

Construction of the slope characteristics requires using a differential curve, such as shown by General Electric[1] for their materials, grades *1* and *2*. The derivation of the slope characteristic for obtaining the curves is described in Appendix II. It is hoped that this paper will encourage thermistor manufacturers to publish the differential curve for other materials or, even better, to publish the resistance-temperature and slope characteristics of Figs. 2 and 3 for the numerous standard thermistor compositions manufactured.

Nomenclature

a = temperature coefficient of thermistor material as a function of temperature
R = resistance of thermistor network as a function of temperature
R' = network resistance as a function of temperature pu (per unit)
R_c = resistance of copper coil
R_o = ratio of thermistor resistance to shunt resistance at 25 C
$R_{oa} = R_o$ for circuit A
$R_{ob} = R_o$ for circuit B
$R_{oc} = R_o$ for circuit C
R_s = shunt resistance
$R_{sa} = R_s$ of circuit A
$R_{sb} = R_s$ of circuit B
$R_{sc} = R_s$ of circuit C
R_t = thermistor resistance as a function of temperature
$R_t' = R_t$ at subnumbered temperature in C
$R_{25} = R_t$ at 25 C

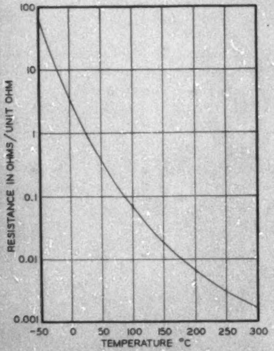


Fig. 8. Resistance ratio temperature characteristic—GE, grade 1

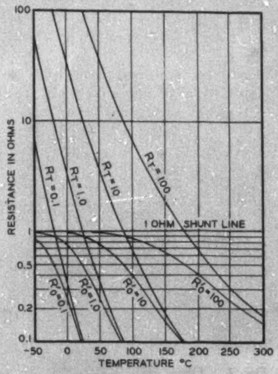


Fig. 9. Resistance versus temperature for various thermistors and their pu networks

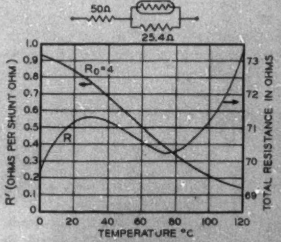


Fig. 10. Graphical illustration, example V

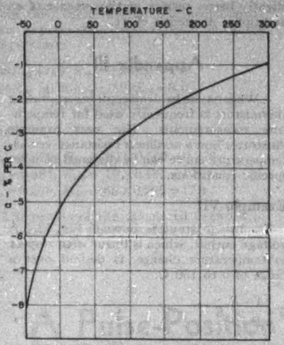


Fig. 11. Temperature coefficient of resistance—GE, grade 1

$$a=\frac{1}{R_t}\times\frac{dR_t}{dT}$$

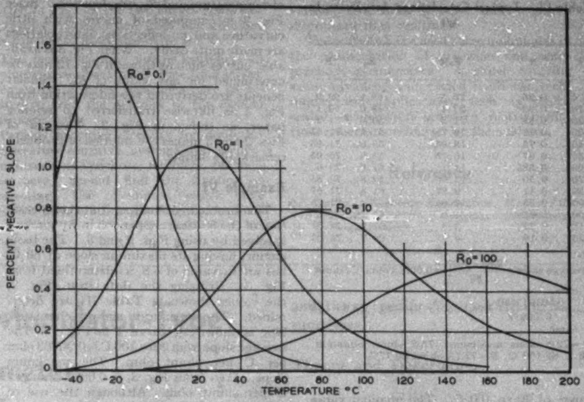


Fig. 12. Slope versus temperature of pu networks—GE, grade 1

$R_{25a}=R_t$ at 25 C, for circuit *A*
$R_{25b}=R_t$ at 25 C, for circuit *B*
$R_{25c}=R_t$ at 25 C, for circuit *C*
R = total resistance as a function of temperature
T = temperature in C
V = terminal voltage
E = source voltage

Appendix I

Resistance-versus-temperature data on various thermistor materials may be found in the literature of thermistor manufacturers. A typical curve of General Electric grade *1* thermistor material, shown in Fig. 8, is rationalized for 1 ohm at 25 C. Fig. 8 may be applied to any thermistor made of this material by multiplying the curve's ordinates by the thermistor's 25 C resistance (R_{25}).

If a thermistor is shunted by a 1-ohm resistor, the combined parallel resistance $R'=R_t/(R_t+1)$, where R_t is the resistance of the thermistor as a function of temperature. Fig. 9 shows a set of grade *1* thermistors having different room-temperature values and the resultant characteristics when they are shunted by 1 ohm. The resultant characteristics asymptotically approach 1 ohm at low temperatures, and the thermistor's characteristics at high temperatures. The family of curves shifts to the higher temperature range as the ratio $R_o=R_{25}/R_s$ increases—where R_s is the shunt resistance.

The advantage of adding the shunt is clearly demonstrated if the characteristics of Fig. 9 are plotted on a linear scale, as shown in Fig. 1. The curves are roughly linear over the temperature range indicated by the solid line part of the curves.

In the application of this basic circuit, the R_o value is often selected from visual inspection of Fig. 1, and the average slope over the desired temperature range determined graphically. The parameters of the pu network ($R_{25}=R_o$) are then multiplied by the ratio of the slope required, divided by the slope of the pu network. The resultant network is then checked for accuracy. If a family of curves, such as in Fig. 1, are not available, then it is usually sufficient to determine R_o by making R_o equal to the reciprocal of the thermistor material's pu resistance at the mid-temperature point. This is based on the assumption that this point is where the thermistor resistance is equal to that of the shunt. As an example, suppose it is desirable to compensate the resistance change of a copper coil over a temperature range of 20 C to 100 C. The coil has a resistance of 50 ohms at 20 C.

Example V

The mid-temperature is 60 C.

$R_{60}=0.25$ for a General Electric grade *1* thermistor.

R_o is $1/R_{60}=4$.

R versus temperature is determined, $R=4R_t'/(4R_t'+1)$.

The average slope between 20 C and 100 C is determined.

R_s and R_{25} are determined.

R' is multiplied by R_s and then added to R_c.

The results are shown in Table I, and plotted in Fig. 10. The graph of Fig. 10 shows that the network has a total resistance of 70.78 ohms ±1.1% over the temperature

Table I. Typical Calculations for Example V

Temperature, C	R_t', Ohms	$\frac{4R_t'}{4R_t'+1}$, Ohms	25.4 R, Ohms	R_c	R
0	3.3	0.929	23.6	46.1	69.7
10	2.05	0.892	22.65	48.0	70.65
20	1.25	0.833	21.15	50.0	71.15
30	0.81	0.764	19.4	51.9	71.3
40	0.53	0.679	17.25	53.9	71.15
50	0.36	0.59	14.98	55.9	70.88
60	0.25	0.50	12.7	57.9	70.6
70	0.175	0.412	10.45	59.8	70.25
80	0.125	0.333	8.45	61.8	70.25
90	0.092	0.269	6.83	63.8	70.65
100	0.068	0.214	5.43	65.7	71.13
110	0.050	0.167	4.24	67.7	71.94
120	0.039	0.135	3.43	69.7	73.13

$$\text{Average slope}=\frac{0.833-0.214}{80}=\frac{0.619}{80}=0.00774\ \text{ohm/C/shunt ohm}$$

$$R_s=\frac{0.00393\times 50}{0.00774}=25.4\ \text{ohms},\quad R_{25}=R_o\times R_s=101.6\ \text{ohms}$$

R = 71.3 maximum; 70.25, minimum; 20 C to 100 C

Table II. Calculations of Slope Characteristic for $R_0=1.0$ of General Electric Grade 1

Temperature, C	Temperature Coefficient of Resistance (a), %/C	Thermistor Value Versus Temperature (R_t)	Per Cent Negative Network Slope, $aR_t/(1+R_t)^2$
−40	7.30	34.0	0.203
−20	6.30	9.8	0.517
0	5.20	3.3	0.928
20	4.52	1.25	1.12
40	4.03	0.53	0.912
60	3.60	0.25	0.576
80	3.20	0.125	0.316
100	2.90	0.068	0.173
120	2.62	0.039	0.095
140	2.38	0.023	0.052
160	2.15	0.0148	0.031
180	1.93	0.0097	0.018
200	1.72	0.0067	0.011

Table III. Typical Calculations for Example VI

T	R′	R_oR'	R_o	R
0	0.95	22.9	46.1	69.0
10	0.92	22.1	48.0	70.1
20	0.88	21.2	50.0	71.2
30	0.82	19.75	51.9	71.65
40	0.75	18.05	53.9	71.95
50	0.67	16.15	55.9	72.05
60	0.585	14.1	57.9	72.0
70	0.50	12.05	59.8	71.85
80	0.41	9.87	61.8	71.67
90	0.33	7.95	63.8	71.75
100	0.27	6.5	65.7	72.3
110	0.22	5.3	67.7	73.0
120	0.18	4.33	69.7	74.23

Average slope $=\dfrac{0.88-0.27}{80}=0.00763$ ohm/C/shunt ohm

$R_s=\dfrac{0.00393\times 50}{0.00817}=24.1$ ohms $R_{sh}=24.1\times 5.8=140$ ohms

R = 72.3 ohms maximum; 71.2 ohms minimum; 20 C to 100 C, R = 71.75 ohms ±0.77%

range of 20 to 100 C. The nominal value of 70.78 ohms is not at 60 C, but at 55 C. This is because the thermistor temperature coefficient is dependent upon temperature. If better performance is required, R_o could be increased slightly and R_s decreased slightly. Since the temperature coefficient is nonlinear with respect to temperature (see Fig. 11), more accurate results will be obtained if R_o is made equal to about 1.2 times the reciprocal of the thermistor material pu resistance at the mid-temperature point.

Appendix II

The use of multiple-series networks requires a more accurate determination of the slopes than can be scaled from the resistance-temperature characteristics in Fig. 1. This requirement is satisfied by differentiating, with respect to temperature, the expression for the resistance of a basic thermistor circuit.

$$R'=\frac{R_t}{1+R_t}$$

and

$$\frac{dR'}{dT}=\frac{dR_t/dT}{(1+R_t)^2}$$

But

$$a=\frac{1}{R_t}\times\frac{dR_t}{dT}\text{ (see Fig. 11)}$$

Then

$$\frac{dR'}{dT}=\frac{aR_t}{(1+R_t)^2}$$

With the use of Figs. 8 and 11 and the foregoing expression, Fig. 12 may be plotted. Typical calculations for $R_o=1$ are shown in Table II for General Electric's grade *1* thermistor material. Simplification and a set of complete general characteristics are obtained by transferring the information from Fig. 12 to semilog paper, as shown in Fig. 3, where the variation in and the percentage of the maximum slope, as a function of temperature and R_o, are depicted. Since Fig. 3 is composed of curves with little curvature and no inflections, interpolations are made quite easily. When Fig. 3 is available, curves such as those in Fig. 12 may be determined for any value of R_o. Similar benefits are derived if the information from Fig. 1 is likewise transferred to semilog paper, as shown in Fig. 2. The use of Figs. 2 and 3 makes the solution of problems remarkably simple.

Example VI

The problem of temperature compensation of the 50-ohm copper coil in Appendix I is solved by using Figs. 2 and 3. The basic circuit, having its maximum slope at 60 C, has an R_o value of 5.8, as determined from Fig. 3. Utilizing the data from Fig. 2, the values shown in Table III are determined. The resultant circuit characteristic is shown by curve 1 in Fig. 7. The average slope from 20 to 100 C is 0.00763 ohm per C per shunt ohm. The maximum slope, taken from Fig. 3, is 0.0087 ohm per C per shunt ohm. Although the use of either slope will give reasonably accurate results, a value halfway between the two is selected.

The results, using this method, compare favorably to those in Appendix I. The resulting circuit has a shunt reduced by 5% and a thermistor increased by 39%. This serves to illustrate the authors' statements concerning permissible network tolerances. In this example, the ease of using Figs. 2 and 3 is shown. Improved accuracy is not demonstrated because the problem's temperature range slightly exceeds the basic circuit's effective range. If the problem were concerned with a range of 60 C instead of 80 C, the technique here would assert its superiority.

Even greater accuracies may be achieved if consideration is given to the asymmetry of the slope-temperature characteristics. Fig. 3 shows that for $R_o=5.8$, the maximum slope is at 60 C, but the slopes reduce to 69% at 100 C, and 55% at 20 C. On the other hand, if $R_o=5$ is selected, the maximum slope occurs at 57 C, but is reduced to 60% at both extremes. This last refinement would be suggested to those who desire slightly better results, at the expense of an additional moment's effort.

Appendix III

A Wheatstone bridge circuit incorporating thermistors is frequently used for temperature measurements. The next example illustrates how a nonlinear resistance-versus-temperature curve can be designed to meet specific conditions.

Example VII

The bridge circuit is shown in Fig. 13. A voltage output, which is linear with respect to temperature change, is desired over a range of 0 to 150 C.

$$V=\frac{1-R}{R+1}E$$

Imposed conditions are: No output voltage at 0 C, or $R=1$ ohm at 0 C.

$R=0.2$ at 150 C

$V=2/3\times E$ at 150 C

$V_o=0$ at 0 C

and

$$V=\frac{TE}{225}=\frac{1-R}{R+1}E$$

$$R=\frac{225-T}{225+T}$$

$R=0.5$ at 75 C

A series network of three thermistors and shunts is used. Fig. 3 provides the three R_o values that have their maximum slopes at 0 C, 75 C, and 150 C. Fig. 2 provides the resistances of these three pu networks which are tabulated as follows:

Circuit	R_o	R′ 0 C	 75 C	 150 C
A	0.4	0.58	0.04	0.00
B	10	0.97	0.60	0.16
C	85	0.99	0.92	0.60

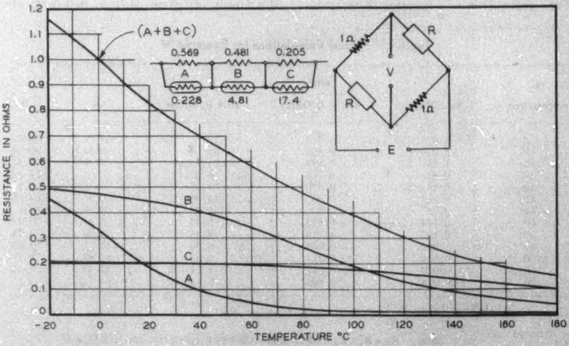


Fig. 13. Three-circuit characteristics for bridge application

From this, these three simultaneous equations express the requirements at three temperatures

$$0.58R_{sa}+0.97R_{sb}+0.99R_{sc}=1.0$$
$$0.04R_{sa}+0.60R_{sb}+0.92R_{sc}=0.5$$
$$0.16R_{sb}+0.60R_{sc}=0.2$$

Therefore

$$R_{sa}=0.569,\ R_{2sa}=0.4\times0.569=0.228$$
$$R_{sb}=0.481,\ R_{2sb}=10\times0.481=4.81$$
$$R_{sc}=0.205,\ R_{2sc}=85\times0.205=17.4$$

The results of the calculated network fit the required nonlinear curve so well that, if plotted in Fig. 13, the two curves could not be readily distinguished from one another. The calculated maximum error is 2 C.

This technique may not be usable in all nonlinear cases because of slope limitations. Frequently, a series resistance may be required to trim the network to the desired resistance value.

Persons requiring circuitry for nonlinear temperature-sensitive systems and components, such as synchros, resolvers, tachometers, etc., should find the application interesting. Use of shunts, which have positive temperature coefficients, extends the possibilities of application, since reversals, and both positive and negative slopes, are then realizable.

The authors have used these techniques in the compensation of precision and semiprecision tachometers. A paper covering more advanced and varied nonlinear techniques and applications, such as multiscaled temperature-sensitive instrumentation, is being considered by the authors.

Reference

1. General Electric Thermistor. *Manual TH-13A*, General Electric Company, Schenectady, N. Y., Aug. 15, 1956.

A Pulse-Position Modulator-Type Magnetometer

S. STRICKER
NONMEMBER AIEE

A. A. WULKAN
NONMEMBER AIEE

A SATURABLE-CORE magnetometer for measuring magnetic fields of the order of a fraction of 1 ma (milliampere) turn per cm (centimeter) is suggested. Its operation is based on a saturable-core element, wherein position-modulated voltage pulses are induced, their shift being linearly related to H_o, the field under investigation. By means of a simple analyzer, information about the unknown field contained in the position-modulated pulses is translated into a d-c voltmeter indication. Linearity, sensitivity, and zero-drift—as influenced by parameter variation and temperature changes—are examined. Agreement is checked between experimental results and those anticipated from theory.

In the past, different types of saturable-core magnetometers have been described.[1,2] Their operation was based on one or two saturable-core elements, which, upon simultaneous subjection to the field under investigation and an external symmetrical a-c magnetizing force, will provide asymmetrical voltage or current pulses. So far, the analysis performed on the waveshape of these pulses was confined to the examination of even harmonic contents, in particular to the generation of second harmonic voltages for the *second harmonic flux-gate magnetometer;* and to comparison of the mean or peak value of successive negative and positive going pulses, for the *differential peak detector magnetometer* and the *bridge-type half-wave, full-wave, and self-saturating-type flux-gate magnetometer.*

The analysis is based on the performance of the saturable-core element as a linear pulse-position modulator. The pulse position with respect to the external symmetrical magnetizing force will vary in accordance with the magnitude and polarity of the unknown field and will be measured by a simple analyzer, converting the position-modulated pulses into width-modulated pulses, whose mean difference is a true indication of the shift and thus of the field under investigation. The basic operation of the analyzer is described, and its lack of zero drift is emphasized, arising from the inherently symmetrical character of the saturable-core element and the balanced nature of the circuitry. The analysis presented here is the result of a semigraphic analytic approach using piecewise linear approximation for the nonlinear flux-field relation. Agreement between experimental results and theory justifies the underlying assumptions.

Theory and Description

The magnetometer in its most general form consists of a saturable-core element, a trigger amplifier, and a bistable multivibrator, as depicted in Fig. 1.

The saturable-core magnetometer element has an open core of highly permeable and easily saturable material, such as, for instance, mumetal, or permalloy. The core is in the form of a thin strip or rod of great-length to small-diameter ratio, surrounded by two separate coils, for output and excitation, respectively. The latter may be supplied from a high-impedance source by a sinusoidal, triangular, or trapezoidal alternating current of such amplitude as will swing the core into saturation twice during each cycle. Consequently, the emf (electromotive force), induced in the output coil because of the symmetrical time-variation of the flux in its core, will be a series of alternatively positive and negative going pulses π radians apart. Upon application of the unknown polarizing field H_o, the symmetry of the B-H loop (hysteresis loop) is upset and asymmetrical flux swings result. The transition of the flux wave from $-B_s$ to B_s will be advanced in time, while that from B_s to $-B_s$ will be retarded, or vice versa, resulting in corresponding voltage pulse shifts.

In order to set up useful mathematical equations for the operation of the magnetometer element, piecewise linear approximation is used for the nonlinear flux-field relation, as depicted in Fig. 2, permitting employment of the assumptions

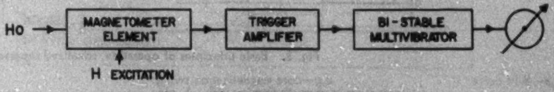


Fig. 1. Block diagram of magnetometer

Paper 61-107, recommended by the AIEE Special Instruments and Auxiliary Apparatus Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted July 26, 1960; made available for printing December 13, 1960.

S. Stricker is with Israel Institute of Technology, Haifa, Israel, and A. A. Wulkan is with the Ministry of Defence, Israel.

This paper is a preliminary report on work carried out under auspices of the faculty of electrical engineering at the Israel Institute of Technology and the Scientific Department, Israel Ministry of Defence.

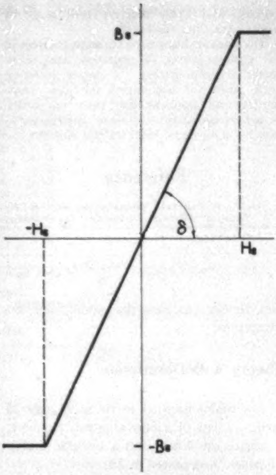


Fig. 2. Idealized magnetization curve

$$\mu = dB/dH = tg\delta = B_s/H_s \qquad -B_s<B<B_s \tag{1}$$

$$\mu = 0 \qquad -H_s>H>H_s \tag{2}$$

$$H_c = 0 \tag{3}$$

zero-hysteresis and eddy-current losses.

Operation Without Signal $(H_o=0)$

Assuming sinusoidal excitation

$$H_{(t)} = H_m \sin \omega t \tag{4}$$

flux swings result in the saturable-core element

$$\phi_{(t)} = \mu A H_m \sin \omega t \qquad -H_s<H_{(t)}<H_s \tag{5}$$

$$\phi_{(t)} = \pm A B_s \qquad -H_s>H_{(t)}>H_s \tag{6}$$

As depicted in Fig. 3, showing a geometrical interpretation of the basic magnetometer operation, the exit from the lower saturation region $(-B_s)$ occurs at

$$\omega t = -\alpha'$$

$$\mu H_m \sin(-\alpha') = -B_s \tag{7}$$

$$\alpha' = \sin^{-1}(B_s/\mu H_m) \tag{8}$$

Entry into the upper saturation region (B_s) occurs at

$$\omega t = \alpha''$$

$$\mu H_m \sin \alpha'' = B_s \tag{9}$$

$$\alpha'' = \sin^{-1}(B_s/\mu H_m) \tag{10}$$

$$|\alpha'| = |\alpha''| = \alpha \tag{11}$$

The voltage across the output coil of the saturable-core element is found from the flux-time derivative therein

$$V_{(t)} = N\frac{d\phi}{dt} \tag{12}$$

$$V_{(t)} = \omega N A \mu H_m \cos \omega t = \omega N \phi_m \cos \omega t \qquad -H_s<H_{(t)}<H_s \tag{13}$$

$$V_{(t)} = 0 \qquad -H_s>H_{(t)}>H_s \tag{14}$$

The coil voltage has the form of symmetrical pulses satisfying the condition $U(\omega t) = -U(\pi+\omega t)$. It consists of a series of positive pulses and a similar one of negative pulses shifted π radians.

In regions

$$-\alpha+2n\pi<\omega t<\alpha+2n\pi \tag{15}$$

$$V_{(t)} = N\omega\phi \cos \omega t \tag{16}$$

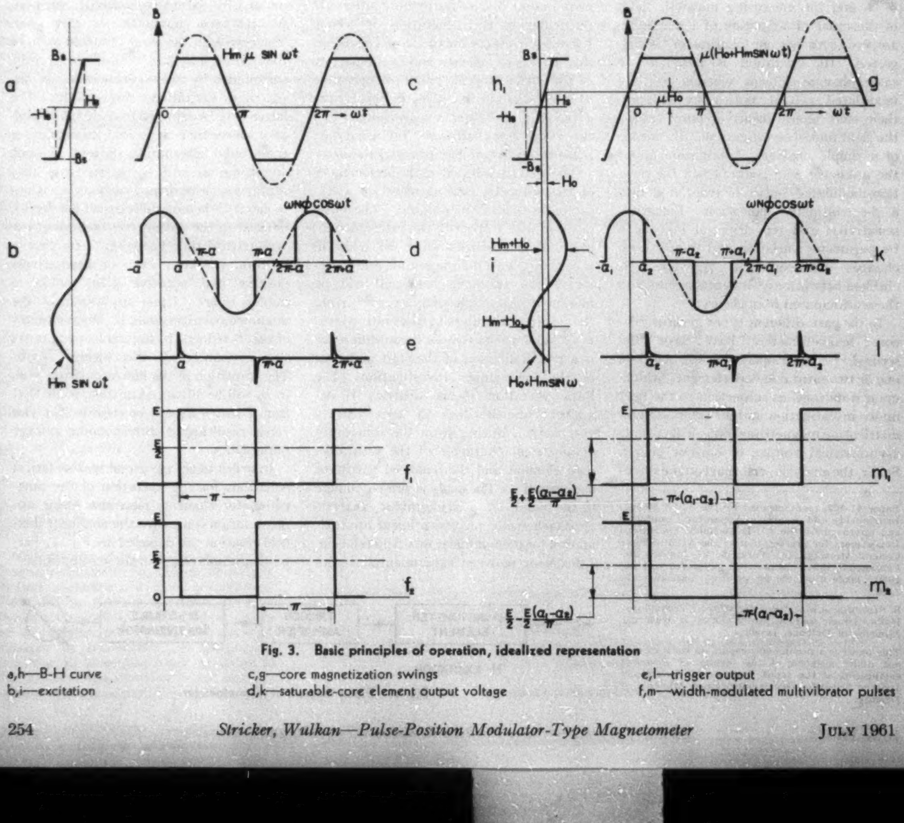

a
b
c
d
e
f1
f2
g
h1
i
k
l
m1
m2
B
Bs
Hs
-Hs
-Bs
Hm μ SIN ωt
0
π
2π → ωt
ωNφcosωt
-α
α
π-α
π+α
2π-α
2π+α
Hm SIN ωt
E
E/2
π
μ(Ho+HmSIN ωt)
μHo
Ho
Hm+Ho
Hm-Ho
Ho+HmSIN ω
-α1
α2
π-α2
π+α1
2π-α1
2π+α2
E/2+E/2(α1-α2)/π
π+(α1-α2)
E/2-E/2(α1-α2)/π
π-(α1-α2)

Fig. 3. Basic principles of operation, idealized representation

a,h—B-H curve
b,i—excitation
c,g—core magnetization swings
d,k—saturable-core element output voltage
e,l—trigger output
f,m—width-modulated multivibrator pulses

In regions

$$-\alpha+(2n+1)\pi<\omega t<\alpha+(2n+1)\pi \qquad (17)$$

$$V_{(t)}=-N\omega\phi\cos\omega t \qquad (18)$$

Successive leading or trailing edges, being π radians apart, are filtered out by a pulse-shaping network and trigger amplifier, and are applied to a bistable multivibrator. The latter is alternately switched ON-OFF with each incoming trigger, resulting in a rectangular wave with identical positive and negative half-cycles about mean value $E/2$ and of magnitude $E/2$. A d-c voltmeter connected between the collectors of the multivibrator will indicate the difference of their mean voltages being equal to zero.

$$V_{\text{out}}=1/2\pi\int_{\alpha}^{\pi+\alpha}Ed(\omega t)-1/2\pi\int_{\pi+\alpha}^{2\pi+\alpha}Ed(\omega t)=0 \qquad (19)$$

The zero of this system is independent of the pulse amplitude; it is not affected by a constant phase-shift as will result from magnetization loop width consideration, or any constant linear delay introduced by the amplifier or multivibrator, as long as symmetry prevails. This applies equally well to changes in amplitude and frequency of the excitation, as long as its zero-axis symmetry persists, and to temperature variations within the range of normal ambient temperatures, merely changing the hysteresis loop without affecting symmetry.

Operation With Signal

With application of the polarizing field ($H_o\pm 0$, $-H_s<H_o<H_s$), the flux-swings are asymmetrically distorted, the transitions being alternately advanced and retarded.

Inspection of Fig. 3 reveals that transitions will take place wherever equation 20 is satisfied

$$\mu(H_o+H_m\sin\omega t)=\pm B_s \qquad (20)$$

Exit from the lower saturation region $-B_s$ occurs at

$$\omega t=-\alpha_1$$

$$\alpha_1=\sin^{-1}\frac{B_s-\mu H_o}{\mu H_m} \qquad (21)$$

Entry into the upper saturation region B_s occurs at

$$\omega t=\alpha_2$$

$$\alpha_2=\sin^{-1}\frac{B_s+\mu H_o}{\mu H_m} \qquad (22)$$

The resultant coil voltage consists again of a series of pulses, but in the present case the condition of symmetry is $U(\omega t)=-U(\pi-\omega t)$, as depicted in Fig. 3.

At the angles $-\alpha_1+2n\pi<\omega t<\alpha_2+2n\pi$

$$V_{(t)}=N\omega\phi_m\cos\omega t \qquad (23)$$

At the angles $-\alpha_2+(2n+1)\pi<\omega t<\alpha_1+(2n+1)\pi$

$$V_{(t)}=-N\omega\phi_m\cos\omega t \qquad (24)$$

Successive leading or trailing edges of the voltage pulses are no longer equidistant, but alternate between amounts $\pi+(\alpha_1-\alpha_2)$ radians and $\pi-(\alpha_1-\alpha_2)$ radians, their difference being equal to $2(\alpha_1-\alpha_2)$ radians.

With the aid of equations 21 and 22, and by introducing $B_s=\mu H_s$, equation 25 can be written

$$\alpha_1-\alpha_2=\sin^{-1}\frac{H_s-H_o}{H_m}-\sin^{-1}\frac{H_s+H_o}{H_m} \qquad (25)$$

Using approximations stipulated in the Appendix, a linear relationship is established between amount $\alpha_1-\alpha_2$ radians and the unknown field H_o.

$$\alpha_1-\alpha_2\approx 2H_o/H_m[1-(H_s/H_m)^2]^{-1/2} \qquad (26)$$

Triggers derived from the leading or trailing edges of the voltage pulses will cause, at the collectors of the multivibrator, rectangular pulses of opposite polarity, exhibiting dissimilar positive and negative pulse duration. The difference of their mean values, being linearly related to amount $\alpha_1-\alpha_2$ and consequently to the unknown field H_o, is found from equation 27

$$V_{\text{out}}=1/2\pi\int_0^{\pi+(\alpha_1-\alpha_2)}Ed(\omega t)-1/2\pi\int_0^{\pi-(\alpha_1-\alpha_2)}Ed(\omega t)=E/\pi(\alpha_1-\alpha_2) \qquad (27)$$

The voltage, being measured by the d-c voltmeter connected between the collectors of the multivibrator, is a true indication of the unknown field H_o.

$$V_{\text{out}}\approx H_o\frac{2E}{\pi H_m}[1-(H_s/H_m)^2]^{-1/2} \qquad (28)$$

The magnetometer sensitivity, defined by dV/dH_o, can be increased by enlarging the supply voltage E and decreasing the excitation H_m.

$$dV/dH_o\approx\frac{2E}{(H_m^2-H_s^2)^{1/2}\pi} \qquad (29)$$

The practical limits to the magnitude of E are imposed by the circuitry employed. The minimum permissible excitation is determined by the magnetometer operation, requiring the excitation field to swing the core into saturation twice during each cycle.

$$H_{m_{\min}}\geq H_s+|H_o|$$

Out of technical considerations, a some-

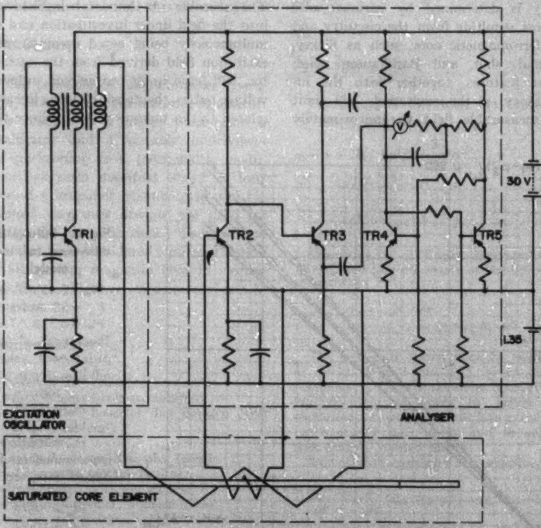


Fig. 4. Basic circuit diagram of magnetometer

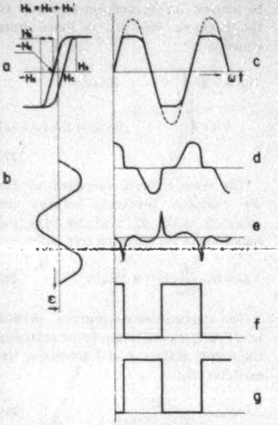


Fig. 5. Basic principles of operation, actual curves from oscillograms for operation without signal

a. B-H loop; b. excitation; c. flux swings; d. coil voltage; e. trigger pulses; f,g. multivibrator output

what larger excitation has to be used to achieve well-defined coil voltage pulses, exhibiting well-defined trailing edges from which the triggers are derived.

The theoretical magnetometer sensitivity is determined by various noise sources resulting from the circuitry and the ferromagnetic core, such as flicker, thermal, shot, and Barkhausen noise. These features, together with the inconstancy of the excitation, will limit the measurable field to approximately 10^{-5} AT/cm (ampere turns per centimeter).

Experimental Results

A preliminary model of a portable magnetometer, incorporating transistorized circuitry, was built along these lines and tested; then experimental results were compared to theoretical ones.

The core used in the experiment was a mumetal strip 2 mm (millimeters) wide, 0.075 mm high, and 12.0 cm long—equivalent diameter being 0.0436 cm ($l/d = 275$).

After heat treatment, the core received a 500-turn output winding at its center and was inserted in a solenoidal excitation coil, which covered its whole length. The rod permeability and the field required for saturation were found by recording the hysteresis loop at the frequency of interest and applying a straight line approximation μ rod = 16,000; $H_s = H_s' + H_c = 0.318$ AT/cm.[3]

Appropriate transistorized circuitry was built, being fed from a 30-volt *B*-battery (one which furnishes direct voltages to plate and electrodes of vacuum tubes) and 1.35-volt mercury cell, the basic circuit diagram being depicted in Fig. 4.[4]

The excitation oscillator *TR1*, employing inductive feedback, operates at 50 cps (cycles per second) in order to enable mains' operation without any alterations. The saturable core element, being inserted into the field under investigation and simultaneously being acted upon by the excitation field derived from the oscillator, will have in its output coil induced voltage pulses, their position being linearly related to the unknown field. Since the actual shape of the hysteresis loop has gradual but dissimilar approach and departure from the saturation region, the voltage pulses will exhibit a gradually rising leading edge but a sharply defined trailing edge, as depicted in Fig. 5.

Table I

Per Cent of Full Scale for H_0	mAT/cm Limit
0.1	10
0.8	20
2.0	50

To secure trigger pulses corresponding to the trailing edges mentioned before, the output coil voltage is amplified by a 2-stage linear amplifier, comprising a grounded emitter stage, *TR3*. Subsequently, the outputs from collector and emitter are differentiated by two identical but separate resistance-capacity differentiating networks, their time constant being of such a value as to switch the bistable multivibrator *TR4*, *TR5* from one steady state to the other, with each incoming trigger which corresponds to the pulse-trailing edge. Although triggering of one multivibrator collector only would have sufficed to insure its proper operation, balanced operation dictated symmetrical double-line triggering. The output voltmeter of a sensitivity of 30 mv (millivolts) per scale division (full scale deflection 3 volts) and 10 kilohms/volt is connected between the collectors of the multivibrator. Shunting condenser *C* serves to by-pass the alternating component passing through it.

With this circuitry, the sensitivity was reduced to 0.5 mAT/cm per scale division. Because pulse-width measuring-equipment, having a resolution of better than 10 microseconds, was not available, data for increased sensitivity could not be collected.

Measurements of pulse width with the above-mentioned equipment checked the pulse position as a function of the polarizing field. With each run, one parameter was varied. Some indicative data appear in Fig. 6 and in the subsequent notes.

With the present equipment, zero-drift was undetectable in temperature runs ranging from 0 C (degrees centigrade) to 55 C. A deviation of less than 0.1% in the scale conversion factor resulted.

Frequency variation of 5% was undetectable, and 5% variation of excitation amplitude affected only the scale conversion factor, as anticipated from theory. The linearity was found adequate, as shown in Table I.

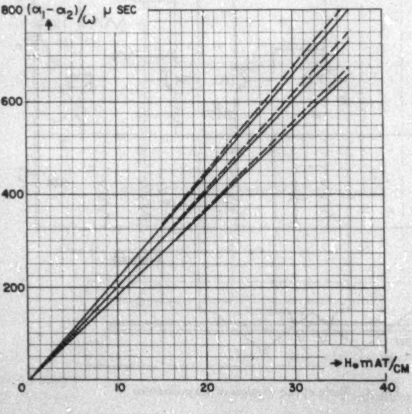


Fig. 6. Pulse-width difference for built model

$H_m = 0.45$ AT/cm
$f = 50$ cycles
$H_s = 0.318$ AT/cm
Theoretical (approximated) values, full lines, top to bottom:
$H_m =$ 95%
$H_m =$ 100%
$H_m =$ 105%
Experimental results, dashed lines, top to bottom:
$H_m =$ 95%
$H_m =$ 100%
$H_m =$ 105%

Table II. Comparison of Types

	Second-Harmonic Magnetometer	Position-Modulator Magnetometer
Number of reactors	1–2	1
Analyzer	amplifier, phase-sensitive demodulator	amplifier, bistable multivibrator
Amplifier	feedback amplifier of constant gain	simple amplifier
Auxiliary circuits	second-harmonic reference voltage generator	none
Critical adjustments	1. 2-core matching 2. balancing of phase demodulator 3. phasing of reference voltage	balancing of bistable multivibrator
System zero	excellent	excellent
Magnitude of magnetometer output voltage	$V_{\text{out}} \approx c_1 G f \frac{H_0}{H_m}\left[1-\left(\frac{H_s}{H_m}\right)^2\right]^{-1/2}$	$V_{\text{out}} \approx c_2 E \frac{H_0}{H_m}\left[1-\left(\frac{H_s}{H_m}\right)_2\right]^{-1/2}$
Dependence on excitation amplitude (H_m)	yes	yes
Dependence on excitation frequency (f)	yes	none
Dependence on amplifier gain (G)	yes	none
Dependence on magnitude of d-c supply voltage (E)	none	yes

Comparison of Magnetometers

Comparison between the position-modulator-type magnetometer and the second-harmonic magnetometer will now be made. Merits of the position-modulator type, its operation, and associated circuitry may be judged by its comparison with the widely used second-harmonic magnetometer, which is generally acceptable for geological survey and submarine detection.[5]

Both types perform an analysis on the waveshape of the output voltage from a saturable-core reactor, which, upon application of the unknown field, becomes asymmetrically distorted.

The second-harmonic magnetometer is essentially a modulator converting the unknown magnetic field into an alternating current wave, the second-harmonic component of which is related to the field under investigation. The analyzer in its basic form consists of a selective amplifier and a demodulator. Because of the presence of large odd-harmonic components, intelligent schemes are required to obviate the use of complex circuitry wherein lies one of its difficulties. Suppression of unwanted frequency components is achieved by the use of two matched reactors with series in-phase excitation and series opposingly connected output coils, or vice versa, and the use of a synchronous demodulator. Actually a phase-sensitive demodulator is used, as magnitude and polarity of the unknown field are investigated. The term "synchronous demodulator" is employed to point out responsiveness to a single-frequency component only.

After the second-harmonic magnetometer is outlined in its basic form, characteristics of both types are projected in Table II.

Conclusions

The principle of magnetometer operation, based on the performance of a saturable reactor as a linear pulse-position modulator and a bistable multivibrator performing as a pulse-width modulator, suggests manifold uses. A low-level d-c amplifier operating in the explained way may obviate the need for costly d-c amplification. The method has also been considered for investigation of alternating magnetic field in electric machines. Further experiments are in progress.

Nomenclature

A = area of core cross section (cm)
B = flux density (volt seconds/cm)
B_s = saturation level of flux density (volt seconds/cm^2)
c = constant
d = equivalent core diameter (mm)
E = d-c supply voltage (volts)
f = frequency (cycles per second)
G = gain of amplifier, dimensionless
H = magnetizing force (ampere turns/cm)
H_o = magnetizing force to be measured (ampere turns/cm)
H_t = coercive force (ampere turns/cm)
H_m = peak value of excitation magnetizing force (ampere turns/cm)
H_s = saturation value of magnetizing force (ampere turns/cm)
1 = core length (cm)
N = number of turns of output coil (turns)
t = general time variable (seconds)
V = alternating voltage (volts)
α = angle (radians)
$\mu = \mu_0$ times relative permeability
μ_{rod} = relative rod permeability of the saturable-core element
Φ = flux (volts seconds)
ϕ_m = peak value of flux (volts seconds)
ω = radial frequency (radians/second)
$\mu_o = 4\pi 10^{-9}$ (volts seconds/ampere turns/cm), omitted for simplicity from all formulas containing μ

Appendix

Linear approximation of relationship between the unknown field H_o and the time-angle difference $\alpha_2-\alpha_1$ is given here. Introduction of $H_s/H_m=a$, $H_o/H_m=b$ into equation 25 results in

$$\alpha_2-\alpha_1=\sin^{-1}(a+b)-\sin^{-1}(a-b)$$

$$\sin^{-1}(a+b)=(a+b)+1/3(a+b)^3+3/40(a+b)^5+\ldots$$

$$\sin^{-1}(a-b)=(a-b)+1/3(a-b)^3+3/40(a-b)^5+\ldots$$

The difference of the linear members on the right side is equal to $2b$. The difference of the second and third members is found after developing the third power, the result being, respectively,

$$a^2b+\frac{1}{3}b^3$$

and

$$\frac{3}{4}a^4b+\frac{3}{2}a^2b^3+\frac{6}{40}b^5$$

Continuing in this way, the expression for $\alpha_2-\alpha_1$ radians is

$$\alpha_2-\alpha_1=2b(1+a^2/2+\frac{3}{8}a^4+\ldots)+ b^3/3(1+\frac{9}{2}a^2+\ldots)+ \frac{3}{20}b^5(1+\frac{25}{2}a^2+\ldots)$$

Discovery of a linear dependence on H_o/H_m will follow a study of the first term to the right

$$\alpha_2-\alpha_1\approx 2b(1-a^2)^{-1/2} \approx 2H_o/H_m[1-(H_s/H_m)^2]^{-1/2}$$

The nonlinear terms may be neglected as $b=10^{-1}\div 10^{-3}$ and $a<1$, based on the reasoning: H_o lies in the range of mAT/cm. $H_m>H_s+|H_o|$ in order to insure the excitation to swing the core into saturation twice during each cycle, and to achieve well-defined voltage pulses.

The reactor sensitivity can be defined by

$$\frac{d(\alpha_2-\alpha_1)}{db}=\frac{d}{db}[\sin^{-1}(a+b)-\sin^{-1}(a-b)] =[1-(a+b)^2]^{-1/2}+[1-(a-b)^2]^{-1/2} =2(1-a^2)^{-1/2}-2[1-(1-b^2)^{-1/2}]$$

For all practical purposes, b^2 may be neglected with respect to unity, resulting in

$$\frac{d(\alpha_2-\alpha_1)}{db} \approx 2(1-a^2)^{-1/2}$$

The over-all magnetometer sensitivity may be defined by dV/dH_o. See equation 29 showing linear dependence on E and dependence on $(1-a^2)^{-1/2}$

$$\frac{dV}{dH_o} = H_m\frac{dV}{db} = \frac{1}{\pi}H_mE\frac{d(\alpha_2-\alpha_1)}{db} \qquad (29)$$

$$= \frac{2}{\pi}EH_m(1-a^2)^{-1/2}$$

References

1. Magnetic-Amplifier Circuits (book), W. A. Geyger. McGraw-Hill Book Company, Inc., New York, N. Y., 1957.

2. Instrumentation for Geophysical Exploration. *Review of Scientific Instruments*, New York, N. Y., vol. 24, no. 4, Apr. 1953.

3. Ferromagnetism (book), R. M. Bozorth. D. Van Nostrand Company, Inc., Princeton, N. J., 1951, 1959, chap. 19.

4. Handbook of Semiconductor Electronics (book), L. P. Hunter. McGraw-Hill Book Company, Inc., 1956.

5. Air-Borne Magnetometers for Search and Survey, E. P. Felch, W. J. Means, T. Slonczewski, L. G. Parratt, L. H. Rumbaugh, A. J. Tickner. *AIEE Transactions*, vol. 66, 1947, pp. 641–51.

Taut Band Suspension Switchboard Instruments

J. C. NYCZ
ASSOCIATE MEMBER AIEE

V. S. THOMANDER
MEMBER AIEE

R. C. MacINDOE
ASSOCIATE MEMBER AIEE

D. F. ROERTY
ASSOCIATE MEMBER AIEE

THE FIRST taut band suspension instrument completely developed, designed, and manufactured in the United States was introduced by the Westinghouse Electric Corporation in June 1958. This instrument, a switchboard type with a d-c moving-coil mechanism and circular scale dial, described in detail in a previous paper,[1] has been in full production for commercial and military applications for more than a year.

The many outstanding advantages and features inherent in a taut band suspended mechanism are now confirmed by more than $2^1/_2$ years' actual field experience, during which additional noteworthy features became evident. Major gains realized through application of taut band suspension principles to electric-measuring-instrument mechanisms are here summarized.

1. No bearing friction: There is inherent friction between jewel and pivot bearings. But the taut band suspension design rules out movement between any two bearing parts that are in contact with each other, and hence eliminates friction.

2. Perfect repeatability of instrument readings: The rolling friction inherent even in new and perfect jewel and pivot assemblies can be detected by noting the difference between upscale and downscale readings or the reading before and after tapping an instrument. With the use of the new ribbon suspension, the repeatability of the instrument is so perfect that an instrument need never be tapped to obtain a more accurate reading.

3. Immunity to vibration: Absence of possible deterioration in the bearing system precludes adverse effects of vibration. Taut band suspension mechanisms have been successfully applied on railway locomotives, naval vessels, and various industrial applications having severe vibration problems.

4. Immunity to shock: The taut band suspension designs include bumpers on the moving elements and bumper stops on adjacent stationary members to relieve the suspension system of all excessive stresses.

5. Reduced damage in shipment and lower installation cost: Items 3 and 4 make these advantages self-evident. Shipping of apparatus with instruments factory-mounted on switchboards permits complete factory check-out and reduces installation cost. The damage to pivots and jewels normally encountered in transportation is nonexistent in taut band suspension instruments.

6. Reduction of field maintenance: The jewels and pivots of electrical measuring instruments are generally recognized as the parts most vulnerable to mechanical damage and wear. In replacing them by the practically indestructible taut band suspension, that area of field service is eliminated.

7. Increased instrument life: It was noted in the 1959 paper[1] that no wear was found in taut band suspension mechanisms after 8×10^6 zero to full-scale excursions of the pointer. Continuation of original tests now increase that figure to 18×10^6 operations, and the original instruments are still on life test without exhibiting any significant change in performance. Conventional pivot and jewel bearing systems require servicing at $3/4\times10^6$ operations.

8. Improved mechanism sensitivities: The energy requirements of pivot and jewel bearing mechanisms are generally determined by the degree to which the effects of bearing friction can be tolerated. In taut band suspension mechanisms, the criterion is generally the speed with which the pointer is expected to respond to changes in the measured electrical quantities. By careful design, particularly of the instrument moving elements, it is possible to improve sensitivities up to several hundred per cent over the pivot and jewel designs. By the same token, the energy loss (burden) of these instruments is radically reduced, and in some cases to a point where low-energy transducer outputs can be measured directly without using amplifiers as was formerly required.

Original development work on the first instrument was inspired by confidence in attaining the then anticipated benefits. Their full realization left no doubt regarding the logic and need of incorporating the taut band suspension principle to all switchboard instruments as rapidly as possible. This having been done, this paper now treats on the design, development, and performance of three taut band suspension mechanisms:

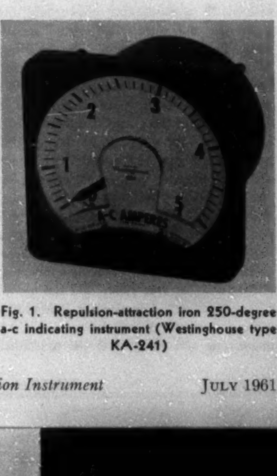


Fig. 1. Repulsion-attraction iron 250-degree a-c indicating instrument (Westinghouse type KA-241)

Paper 61-65, recommended by the AIEE Indicating and Integrating Instruments Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 31, 1960; made available for printing January 10, 1961.

J. C. Nycz, V. S. Thomander, R. C. MacIndoe, and D. F. Roerty all are with Westinghouse Electric Corporation, Newark, N. J.

1. An a-c mechanism for 250-degree scale switchboard instruments for measuring voltage and current. This is the a-c counterpart of the d-c mechanism described previously.[1]
2. A d-c mechanism for 100-degree scale switchboard instruments for measuring voltage and current.
3. An a-c counterpart to item 2.

Significantly, in each of the new mechanisms, taut band suspension details or parts, except for the ribbon itself, are identical to the original.[1] The suspension ribbon differs only as to thickness, which dimension is determined by the response-time requirements and inertia characteristics of each individual mechanism's moving element.

The 250-Degree A-C Switchboard Instrument (Fig. 1)

Repulsion-Attraction Vanes

As may be seen from Fig. 1, the 250-degree a-c instrument is similar in appearance to the d-c type. Both mechanisms are housed in a shorter case than pivot and jewel designs and can be distinguished by title markings on the dials. The a-c mechanism has been completely redesigned in order to use taut band suspensions. This line indicates a-c volts, amperes, and milliamperes.

The mechanism operates on the repulsion-attraction iron-vane principle. The initial deflection of the moving element is caused by instantaneously induced electromagnetic poles of similar polarity on the corresponding ends of a stationary repulsion vane and the moving vane. The upper part of the scale or pointer travel is caused by instantaneously induced electromagnetic poles of opposite polarity on the corresponding ends of two attraction vanes and the moving vane; see Fig. 2 for a schematic or a developed view showing the relationship of these torque-producing vanes. Fig. 3 shows the vanes assembled in a cutaway drawing of the instrument mechanism.

With reference to the developed view of the vanes (Fig. 2) as the pointer travels up scale, or the moving iron travels from left to right in the developed view, the spacing between the ends of the moving vane and the polarized edges of the repulsion vane increases. At the same time, the spacing between the ends of the moving vane and the polarized edges of the attraction vanes decreases. As a result of this changing distance between the ends of the vane combination, the magnetic repulsion forces diminish with the increasing distance, and the magnetic attraction forces increase with the decreasing distances. The shape of the edges of each of the various vanes are developed to blend repulsion and attraction forces against the opposing twist force of the suspension bands to give substantially a uniform scale distribution.

The measured voltage or current applied to a solenoid-type coil polarizes the vanes, which in turn produce pointer indication. The two adjustable attraction iron vanes are concentric with the moving and repulsion vanes, and are assembled against the shield at either end of the coil. The shield is a closed iron circuit of thin steel strips for a low magnetic reluctance return path for the coil field and for shielding the vanes from extraneous fields.

To calibrate the instrument, the screws holding the attraction vanes against the shield are loosened and the vanes turned until calibration is obtained. In practice, the rear attraction vane is set in a predetermined optimum position, and the front one then acts as a trimmer to obtain the final calibration. This adjustment can be made from the front of the instrument with the cover removed, even when mounted on a switchboard.

Vane Assembly for Stability

To obtain the close tolerances necessary to control the iron circuit for consistent calibration and to maintain instrument stability through temperature, humidity, etc., the complete magnetic circuit of the instrument is mounted on a diecast frame. In previous designs the stationary vanes and shields were attached to the plastic coil form. In this new design, each mechanism frame is machined to provide accurate mounting of the stationary vanes and to achieve concentricity with the centering mounts of the suspension assemblies on either end. The element, vanes, and shields minus the coil are attached to the mechanism frame and inspected for proper mechanical and magnetic relationship. The coil form has the sole function of holding the coil winding, which is put on last and is located around the housing of the repulsion vane.

Limits for 250-Degree Instrument

In a previous paper[1] the dimensional limitations of the suspension for a 250-degree instrument were discussed. To eliminate the zero-set error, a limit in band thickness for a given alloy was determined. Having the band thickness, the the bandwidth is obtained from an

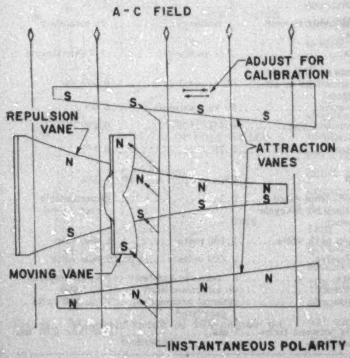


Fig. 2 (left). Developed view of the iron vanes of the 250-degree a-c indicating instrument

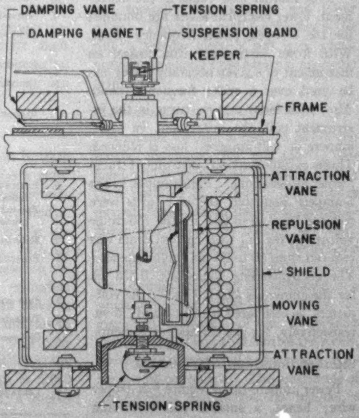


Fig. 3 (right). The 250-degree a-c mechanism

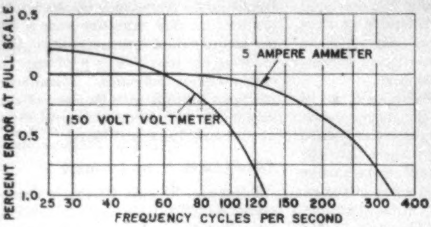


Fig. 4. Frequency response of 250-degree a-c instrument

optimum width-to-thickness ratio. Having both the band thickness and width, the working tension permissible on the given alloy band is evident from tensile tests and reasonable safe loading. When this taut band suspension was applied to the conventional pivot-jewel-type a-c moving element and was placed in a horizontal position, the sag was twice the permissible limits. It was therefore necessary to reduce the weight of the moving element by at least half. To accomplish this, the moving vane thickness and the element shaft cross section were brought to 40% of their former values. To maintain a magnetic density for best instrument performance, the thickness of repulsion and attraction vanes was also cut by similar amounts. This was made possible by the reduced magnetic flux requirement resulting from the lower operating torque, and the lower torque is produced by the limited size of suspension bands. The ampere-turns required are thus only 40% of the ampere-turns required for the corresponding pivot and jewel design.

ELECTRICAL LOSSES

With thinner iron vanes in the instrument, lower electrical losses are obtained for the 60-cycle power-frequency input. With lower instrument iron losses, an instrument of a given accuracy rating can be used over a wider frequency band. Also, with fewer ampere-turns in the instrument coil, the instrument va (volt-ampere or volt-amperes) loss is reduced. This is of importance on instrument transformer applications. For example, in the 5-ampere ammeter, the coil turns were reduced from 50 to 20. In the 150-range voltmeter, the full-scale current drain was reduced from 50 to 20 ma (milliamperes). The corresponding losses are 0.375 va for the ammeters and 3.0 va for the voltmeters, as compared with 1.35 and 7.5 va, respectively, for the former instruments.

SMALLER INSTRUMENT MECHANISM

Fewer ampere-turns require less copper wire; hence, a smaller coil form and shorter mechanism which allows housing in a smaller case. On voltmeters, the external resistor cage for $7^1/_2$ watts was eliminated. The smaller case dissipates the heat from a 3-watt resistor embedded in the base, and will remain within permissible temperature rise limits for good instrument performance.

PERFORMANCE ON FREQUENCY

The standard 150-range voltmeters can be used as normally calibrated on power frequencies from 25 to 100 cycles and the ammeters from 25 to 200 cycles; Fig. 4. For other frequencies, such as 400 cycles, the standard instrument is well within the American Standard[2] requirements of 1/2% accuracy for a ±10% frequency variation. See Table I.

CALIBRATION ADJUSTMENTS

The previous pivot-jewel-type instrument has two adjustments for calibration or recalibration in the field: (1) the turning of the attraction vanes; (2) spring torque adjustment. Employment of taut band suspension eliminates the spring torque adjustment.

The additional adjustment of spring was found to be unnecessary for these reasons:

1. Greater accuracy was possible in assembling the instrument iron circuit to the mechanism frame as previously described.

2. There was increasingly less variation from instrument to instrument when using completely machined single-piece attraction vanes, instead of blanked, formed, and welded vanes.

3. The suspension band torques are more uniform percentage-wise than are the spring torque; this consistency aids in calibration adjustment required. The torque variations are within ±3% with suspension bands.

The changes in calibration during various life tests (the maximum changes were of the order of 1%) such as heat, age, humidity, and vibration, were so small that a slight touch-up of the front attraction vane recalibrated the instrument. Adjustment limits of the top attraction vane alter calibration as much as 10%.

Table I. Performance of 250-Degree A-C Instrument Versus American Standard C39.1 Requirements

Item	Units	American Standard Requirements	Type KA-241 Performance
1. Case, material, finish		molded thermosetting cover and base and steel case-black matte	
2. Scale length	inches	6.8	7.1
3. Normal operating position	mounted on vertical panel	same	same
4. Position influence	60-degree tilt from operating position	1%	1/2%
5. Rated accuracy	per cent	1	1
6. Damping factor	ratio, full scale to overshoot	5 minimum	25 minimum
7. Response time	seconds	2.5 maximum	1.75 maximum
8. Loss	volt-amperes, full scale:		
	milliammeters or ammeters	5 va maximum	0.375 va
	voltmeters	10 va maximum	3.0 va
9. External temperature influence	milliammeters and ammeters	0.5%	0.2%
For ±10 C (degrees centigrade)	voltmeters	0.75	0.5%
10. External field influence	per cent	3%	1.5%
11. Frequency influence	for ±10% from rated frequency for 60-cycle frequency	0.5%	imperceptible (see curve)
12. Safe working ground volts	maximum peak volts	1,100 volts	1,385 volts
13. Dielectric tests	volts effective	2,600 volts	2,600 volts
14. Self-contained ratings	ammeters	3 amperes and less	1 to 30 amperes
	milliameters	no minimum limit	5 to 1,000
	voltmeters	external accessories permitted	10 to 800 volts
15. Leakage current	800 volts between terminals and case	1 ma	less than 0.01 ma

Moving Vane Shape

The moving vane has a cross-sectional area at the center which is twice that of either end. This variation in area gives a uniform magnetic flux density thoughout the vane length. This vane shape improves performance and gives true indications on distorted waveforms of voltage and currents measured.

The moving vane is shaped to wrap around the shaft for two reasons: (1) to give greater strength in high shock; and (2) to give lower inertia for faster response.

The angle of the attraction vanes is made a true thread pitch to permit accurate machining of the parts. The repulsion vane has a curve edge to compensate fully and gives substantially uniform scales and stability for both voltmeters and ammeters, each having similar scale distributions.

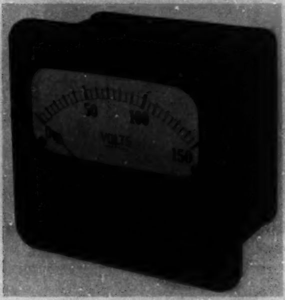


Fig. 5. Core-magnet moving-coil 100-degree d-c indicating instrument with taut band suspension (Westinghouse type KX-251)

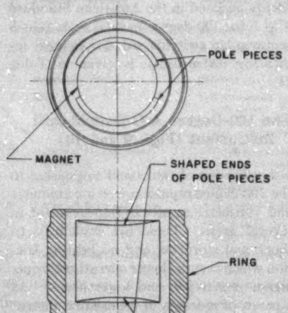


Fig. 6. Magnetic circuit of 100-degree d-c instrument

Response Time

A moving element of low inertia is an important design principle for use with taut band suspensions, as response time for instrument indication can be a limiting factor. Even though the instrument torque is reduced, the response time is essentially the same; it varies from 0.75 second to 1.75 seconds for a damping factor of 100 to 25, respectively.

Performance Tests

Performance tests show that a repulsion-attraction instrument designed to use taut band suspension more than meets the American Standard requirements which apply to this type of instrument.[3] This design also has the regular advantages of taut band suspension as discussed in the first part of this paper, with special mention of lower losses, wider frequency limits, and lighter total instrument weight. The ammeters are 25% lighter and the voltmeters are 33% lighter than are those of the pivot and jewel design.

The 100-Degree D-C Switchboard Instrument (Fig. 5)

The 100-degree d-c instrument employing the taut band suspension bearing system also uses a new core magnet design. Prior to development of the suspension bearing system, core magnet instruments were limited as to obtainable sensitivity. In this design, the combination of the new bearing system (which permits use of lower torques without friction problems), plus a new magnet material, eliminates this sensitivity limitation. As a result, this core magnet, taut band suspension instrument can be made equally as sensitive as the most sensitive pivot-and-jewel large-external-type high-coercive-magnet instrument. With special suspension ribbons and tension springs, quite high sensitivities are possible. A design using extra fine ribbon and light tension springs is now under development.

Permanent Magnet System

In addition to the suspension bearing system, this instrument employs a core-magnet design of novel construction; see Fig. 6. The magnet is a true cylinder, centerless ground for best possible accuracy of air-gap dimensions, with no flats or other shapes normally used to obtain a linear scale. The centerless ground cylindrical magnet has obvious mechanical advantage, plus an increase in the flux density in the air gap, over odd-shaped magnets. The flux density in the air gap of this magnet design is 2,050 gauss. The magnet material, a recent development of the unoriented alnico group, is also advantageous in the mechanical assembly since no effort has to be expended in lining up the direction of orientation. As in all such magnet designs, soft iron pole pieces are used. Linear scale distribution is obtained by using pole pieces of appropriate thickness and specially shaped upper and lower edges. The return path for the flux is the ring, a tubular piece of soft iron, which provides shielding from external magnetic fields.

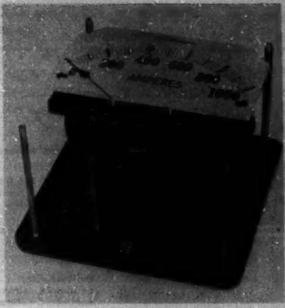


Fig. 7. The 100-degree d-c instrument removed from case

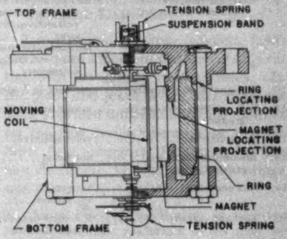


Fig. 8. The 100-degree d-c mechanism

Mechanical Construction

A unique method is used to locate the magnet assembly, ring, and moving element, and to provide concentricity among these parts; Figs. 7 and 8. Two similar diecast frames, one at the top of the mechanism and one at the bottom, are used. The frames, which have two projections on each side to locate and hold the magnet assembly and ring, are machined in a single setup on a turret lathe. Thus, the machined diameters of the projections and the tapped hole that locates the screw, which in turn locates the suspension bearing, are all concentric. The magnitude of the diameter that locates the magnet is

Table II. Performance of 100-Degree D-C Instrument Versus American Standard C39.1 Requirements

Item	Units	American Standard Requirements	Type KX-251 Performance
1. Case, material, finish			metal, black matte
2. Scale length	inches	5	$5^1/_{16}$
3. Normal operating position	mounted on vertical panel	same	same
4. Position influence	60-degree tilt from operating position	1%	0.3%
3. Rated accuracy	per cent	1	1
6. Damping factor	ratio, full scale to overshoot	5 minimum	200–400
7. Response time	seconds	2.5 maximum	0.8
8. Loss	milliammeters	500 mv maximum	50 mv
	ammeters, self-contained	150 mv maximum	50 mv
	ammeters, external-shunt	50 mv ± 1%	50 mv ± 1%
	millivoltmeters, including external-shunt ammeters	25 ma maximum	0.7 ma
	voltmeters	11 ma maximum	1 ma
	kilovoltmeters	1 ma ± 1%	1 ma ± 1%
9. External temperature influence for ±10 C	milliammeters, 10 ma and over; ammeters; millivoltmeters 50 mv and over	1%	0. %
	milliammeter under 10 ma; voltmeters; kilovoltmeters	0.5%	imperceptible
10. External field influence	per cent in 5-oersted field	3	less than 1%
11. Safe working voltage to ground	maximum peak volts	1,100	1,385
12. Dielectric test	volts effective	2,600	2,600
13. Maximum shunt lead resistance	not listed		3.6 ohms
14. Self-contained ratings	ammeters	50 amperes and less	same
	voltmeters	800 volts and less	same
15. Leakage current	800 volts between terminals and case	1 ma	less than 0.01 ma

such that in all cases the magnet has to be forced into position. In order to avoid shaving the aluminum castings when this press fit is made, the edges of the magnet are rounded. This not only locates the magnet concentric to the other parts but also holds the magnet securely in place. The ring is accurately located by machining the diameters involved to close tolerances, and is held mechanically in place by the two screws holding the two castings together.

The self-centering tension spring assembly is exactly the same as that used on the other suspension-type instruments described in this paper. The moving element parts are held concentrically by the accurate assembly fixture, while the ribbon is centrally positioned to these parts by the band guide and anchor.[1]

ADVANTAGES

The advantages of the suspension design have been described earlier. The major advantage of the core-magnet system is the high degree of shielding from external magnetic fields. In this design, a stray field of 500 oersteds will produce less than 1.0% error. Also, since completely symmetrical parts are used, assembly is greatly simplified.

Moreover, the mechanical overload that this mechanism can stand is appreciably greater than that of conventional designs. Sudden overloads of up to 150 times full-scale value can be applied with less than 1% change in calibration. The instrument will stand even greater overloads than this if the activating quantity is applied gradually. Currents of 220 times full-scale can be passed continually through the band and coil of the standard designs without damage.

Further advantages result from the increase in current sensitivity of millivoltmeters from 0.024 amperes for the standard pivot-jewel type to 0.7 ma for the standard taut band suspension design. These advantages are:

1. Shunt leads having a maximum resistance of 3.6 ohms can be used with 50-mv (millivolt) shunts as compared with only 0.15 ohm for the standard pivot and jewel design.

2. Because of the high internal resistance of millivoltmeters, which results from their low current drain, greater variations in lead resistance can be tolerated. A variation of as much as 0.3 ohm in the resistance of the leads will produce an error of less than 0.5%.

The 100-degree d-c switchboard instrument described employs all the major developments of the art up to the present time: the taut band suspension bearing system, a core magnet, and a unique mechanical assembly. Performance of the type *X-251*, as compared with requirements outlined in the American Standard *C39.1* for 100-degree normal-scale 6-inch rectangular switchboard d-c electric indicating instruments, is given in Table II.

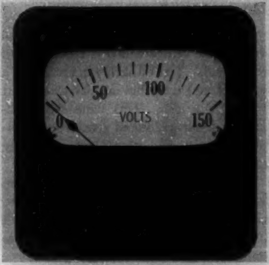


Fig. 9. Repulsion-iron 100-degree a-c indicating instrument with taut band suspension (Westinghouse type KA-251)

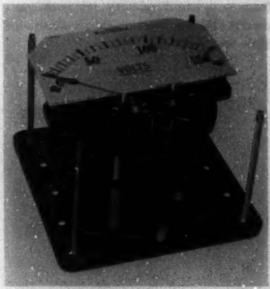


Fig. 10. The 100-degree a-c instrument removed from case

The 100-Degree A-C Switchboard Instrument (Figs. 9 and 10)

Application of taut band suspension to the 100-degree repulsion-iron a-c ammeter and voltmeter results in improvement in several areas. These are immunity to shock and vibration and no bearing friction which permits lower operating torque, higher sensitivity, and lower losses. As a result of redesign of an existing instrument to gain these benefits of the taut band suspension system, additional improvements and innovations were made possible in the areas of: calibration adjustment, frequency range, magnetic damping, elimination of end-play problems, and shielding from stray fields.

Reference to the old design provides a basis for determining advances in the new.

Sensitivity and Loss

Starting with fixed case and pointer size, calculations involving the weight and inertia of the moving element, coupled with the capabilities of taut band suspension, indicated that a full-scale operating torque in the order of 0.5 to 0.7 millimeter-gram should be used. A torque of 0.61 millimeter-gram was chosen since a taut band giving this torque had already been designed for the d-c instrument described. Compared with the old pivot and jewel design, full-scale operating torque was reduced from 1.23 millimeter-grams to 0.61 millimeter-gram. Hypernik moving and stationary vanes were retained but reduced in thickness from 0.020 inch to 0.010 inch. Magnetic flux in the vanes remained at 5,000 gausses, but the ampere-turn requirement now was only 100 compared with 180 ampere-turns for the pivot and jewel design. Coil turns were reduced by the ratio of 100 to 180, keeping the same wire sizes. For equivalent ranges, coil resistance was accordingly reduced to approximately half, and coil inductance to one fourth. Loss values are given in Table III.

Calibration Adjustment

Heretofore, the conventional method of calibrating repulsion-iron instruments was by changing the restraining torque of the instrument. This was usually accomplished by clamping the helical control spring at different points on its outer convolution, thereby changing the length of the spring and hence reducing or increasing restraining torque. In the taut band suspension system, the control element is now a torsion spring whose length cannot be varied in a practical manner.

It is known that developed torque can be changed by moving or translating the stationary vane relative to the moving vane in repulsion instruments. Calibrating may be done in this manner without adjusting a control spring. Experiment showed, however, that care must be taken in choosing the plane of motion of the stationary vane. Arbitrary movement, although changing sensitivity, may have an adverse effect on scale distribution. The method found most suitable, both for maintaining consistent scale distribution and satisfying mechanical considerations, can be described by referring to Fig. 11.

The stationary vane is attached to a movable arm, the pointed projection of which, at its lowest point, is seated in a "V" depression in the frame. This "V" serves as a pivot point for the arm, which is secured at its uppermost curved edge by a spring clamp. By grasping the projection at the upper end of the arm, the arm can be moved in a radial plane close to the plane of the moving vane in its zero position. By thus displacing the stationary vane with respect to the moving vane, sensitivity is changed and calibration obtained. The arm is conveniently located near the front of the instrument and with its cover removed, calibration or adjustment is easily performed. A total adjustment range of 18% is available with this method.

Shielding from Stray Fields and Magnetic Damping

Errors caused by stray magnetic fields can be quite serious in instruments operating with low-density working fields. In the pivot and jewel design, the steel case alone provided sufficient shielding, but not for the new design operating at a much lower field strength. In developing the new shield, Fig. 11, a threefold advantage was taken:

1. An unshielded mechanism is sensitive to any magnetic materials which may be encompassed by the working field. These magnetic paths may be of such extent that they materially affect the ampere-turn

Table III. Performance of 100-Degree A-C Instrument Versus American Standard C39.1 Requirements

Item	Units	American Standard Requirements	Type KA-251 Performance
1. Case, material, finish			metal, black matte
2. Scale length	inches minimum	5	5$^1/_{16}$
3. Normal operating position	mounted on vertical panel	same	same
4. Position influence	60-degree tilt from operating position	1%	0.3%
5. Rated accuracy	per cent	1	1
6. Damping factor	ratio, full scale to overshoot	5 minimum	25 minimum
7. Response time	seconds	2.5 maximum	1.2
8. Loss	self-contained 60-cycle voltmeters	15 va maximum	3 va
	self-contained ammeters and milliammeters	5 va maximum	0.45 va
9. External temperature influence	with 10 C variation from +25 C voltmeters	0.75% maximum	0.05%
	ammeters and milliammeters	0.5% maximum	imperceptible
10. External field influence	in 5-oersted field	3% maximum	0.75%
11. Frequency influence	with ±10% deviation from rated frequency	0.5% maximum	imperceptible
12. Safe working voltage to ground	maximum peak volts	1,100	1,385
13. Dielectric test	volts, live parts to face	2,600	2,600
14. Self-contained ratings	ammeters	50 amperes and less	50 amperes and less
	voltmeters	external accessories permitted for all ranges	10 to 800 volts self-contained
15. Leakage current	800 volts between terminals and case	1 ma	less than 0.01 ma

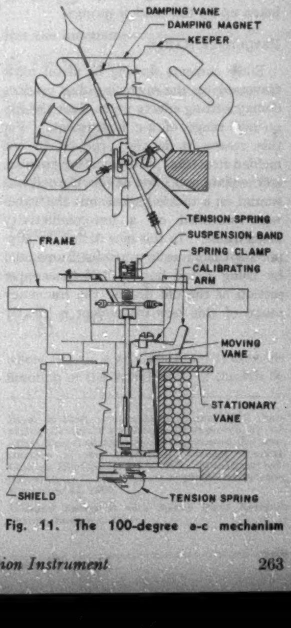


Fig. 11. The 100-degree a-c mechanism

sensitivity. The new shield, Fig. 10, is made of low-loss high-induction hypernik, forming an open-ended box enclosure for the mechanism. The low-reluctance path afforded by this shield carries most of the operating flux and prevents the working field from "seeing" external magnetic materials. This design allows this same mechanism, without change, to be used in other instrument designs wherein cases and mounting structures vary in magnetic path configuration.

2. The effect of external magnetic influences has been reduced to one fourth the American Standard requirement.

3. An eddy-current damping system is used. A magnetic damping system is desirable because of its simplicity of construction and assembly with the added ease of maintaining consistent damping factors. The shield effectively protects against the stray field effect of the damping magnet. No errors were found in tests run with and without damping magnets when shielded in the manner described.

Elimination of End-Play Problems

In order to keep eddy-current losses at a minimum, nonmetallic structural parts are preferable wherever such parts would be embraced by the magnetic field of the coil. Molded members suffering dimensional changes from aging or varying temperature and humidity conditions have been a problem, not only during assembly but after an instrument is in use. Conventional pivot and jewel instruments required close tolerance end-play adjustment of about 0.002 to 0.005 inch. With taut band suspension, any dimensional change is automatically taken up by the tension springs.

Frequency Range

Eddy currents flowing in metal parts traversed by the operating flux produce demagnetizing effects which limit the frequency range of a-c instruments. For this reason it is the practice to use molded structural parts wherever possible. In the pivot and jewel design, the coil was wound on a molded spool, but the vanes were mounted on a low-conductivity brass frame. In this new design, by supporting the vanes in a molded frame, and by using a coil having half the copper section of the old design, the frequency range of this new instrument is greatly improved. The frequency characteristics of the new design are shown in Fig. 12.

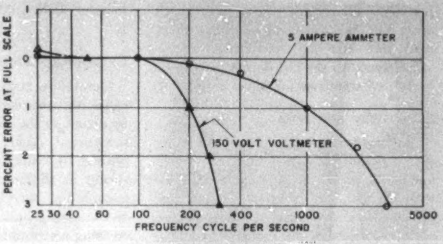


Fig. 12. Frequency response of 100-degree a-c instrument

American Standards Requirements

Gains have been made in most of the areas covered by American Standard *C39.1*. Performance data of the *A-251* line of instruments compared with the American Standards requirement are shown in Table III.

Conclusions

The taut band suspension mechanisms described provide a comprehensive coverage for measuring alternating and direct voltage and current. However, on electric switchboards, measurement of other quantities such as power, reactive power, frequency, power factor, and synchronism is equally important.

In the case of the frequency meter and the watt and var meters, recent developments of appropriate transducers make it possible and desirable to measure frequency and power on the d-c taut band suspension instruments. The development and features of the frequency transducer have already been presented in an AIEE paper.[3] The output of this transducer can be applied to either the 100-degree or 250-degree d-c taut band suspension instrument, or a number of them in series, if indication at more than one point is desirable. The transducer-instrument combination is calibrated on a variable-frequency source and the instrument dial is marked accordingly.

Measuring power on an a-c system is accomplished with a d-c taut band suspension instrument by a transducer employing the Hall effect, described in an unpublished paper by T. Barabutes and W. J. Schmidt. Again, the d-c output of the Hall effect transducer can be applied to either the 100-degree or 250-degree taut band suspension instrument with its scale calibrated in watts.

In the case of a synchroscope, its moving element must be free to rotate unrestrained and with no opposing torque through any number of complete successive revolutions like the armature of a motor. The taut band suspension principle is obviously not applicable to this type of mechanism. Similarly, power factor is measured on the same type of rotating vane mechanism, and it, too, does not employ the taut band suspension principle.

Thus, it is now possible for the first time to obtain switchboard instruments with the taut band suspension principle and its many significant and outstanding advantages in both the 100-degree or 250-degree scale configurations for the measurement of alternating and direct voltage and current, frequency, a-c power, and reactive power.

References

1. Taut Band Suspension for 250-Degree Instruments, V. S. Thomander, R. C. MacIndoe. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, Sept. 1959, pp. 379–84.

2. Electrical Indicating Instruments. *C39.1-1959*, American Standards Association, Inc., New York, N. Y., 1959.

3. Transducer-Type Frequency Meter, U. L. Smith. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, Sept. 1955, pp. 497–500.

E-Algebras in Switching Theory

WARREN SEMON
NONMEMBER AIEE

CERTAIN CLASSES of switching functions, notably the symmetric functions, are realizable by iterative networks.[1-3] It is the purpose of this paper to generalize this concept.

A "switching function of n variables" is defined as a function from the Cartesian product $\Gamma = X_1 \times X_2 \times \ldots \times X_n$ to Z where $X_1, X_2, \ldots, X_n, Z$ are sets consisting of two elements each. The set of all such functions may be identified with the Boolean algebra A of order 2^{2^n}, consisting of the subsets of Γ (which has 2^n elements called "configurations") and the operations of set union, intersection, and complementation. The elements of Γ will sometimes be identified with "configuration numbers," integers i in the interval $0 \leq i \leq 2^n - 1$, the correspondence being the usual one.[4]

E-Algebras

Let N be the set of indices of the argument sets $X_1, X_2, \ldots, X_n$. Thus $N = \{1, 2, \ldots, n\}$. Let $\mathcal{R}$ be an equivalence relation defined on N. A relation R on Γ may be defined as follows:

Let $x, y \epsilon \Gamma$, so that

$$x = (x_1, x_2, \ldots, x_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

Define xRy if and only if there exists a permutation P of the index set N such that if

$$P(1, 2, \ldots, n) = (i_1, i_2, \ldots, i_n)$$

then

$$x_k = y_{i_k}$$

and

$$k \mathcal{R} i_k \qquad k = 1, 2, \ldots, n$$

THEOREM 1

"R is an equivalence relation."

Proof: (1). R is reflexive, since P may be taken as the identity and $k \mathcal{R} k$ for all k. (2). R is symmetric, since the existence of P implies the existence of P^{-1}, and $k \mathcal{R} i_k$ implies $i_k \mathcal{R} k$. (3). R is transitive since $\mathcal{R}$ is transitive, and the existence of P_1 and P_2 implies the existence of their product.

An equivalence relation on a set corresponds biuniquely to a partitioning of the set into collectively exhaustive disjoint subsets, called "equivalence classes." Let it be assumed that $\mathcal{R}$ partitions the n indices of N into r equivalence classes and that R partitions Γ into p classes $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$. The Boolean algebra consisting of the subsets of the set $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_p\}$ has 2^p elements and is isomorphic to the subalgebra of A generated by the p "elementary functions" defined by the subsets Γ_i. Any subalgebra of A induced by an equivalence relation $\mathcal{R}$ on N, extended to R on Γ will be called an "elementary E-algebra." Thus an elementary E-algebra is a subalgebra induced by an equivalence relation.

THEOREM 2

"Each elementary function Γ_i is invariant to any permutation of arguments with indices in a single $\mathcal{R}$-equivalence class."

Proof: Let

$$x = (x_1, x_2, \ldots x_n)$$

be an arbitrary configuration belonging to Γ_i, and let Q be any permutation of the indices in a single $\mathcal{R}$-equivalence class. Then

$$Q(x) = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$$

and by hypothesis $k \mathcal{R} i_k$ for $k = 1(1)n$. To show that $Q(x) \epsilon \Gamma_i$, choose P as Q^{-1}. Then $P(Q(x)) = x$ and this choice of P clearly satisfies the condition of the definition.

It follows at once from Theorem 2 that the equivalence classes into which N is divided may be thought of as defining permutations of the argument sets to which the functions of the elementary E-algebra are invariant. Hence the equivalence classes defined by $\mathcal{R}$ correspond to a subgroup H of the group G of all permutations of the arguments. Furthermore, the set of all functions in the elementary E-algebra contains those functions which belong to H, i.e., are invariant to all permutations of H. If transformations of argument complementation are now included, then a subgroup $c^{-1}Hc$ conjugate to H, where c is an argument complementation, induces a subalgebra of A which is isomorphic to that induced by H.

An example of an elementary E-algebra is the set S of symmetric functions. It is clear that S comprises the elements of a subalgebra since the sums, products, and complements of symmetric functions are again symmetric functions. The relation $\mathcal{R}$ on N is the universal relation, i.e., $i \mathcal{R} j$ for all i, j in N, whereas the relation R on Γ is the "k out of n" relation. Hence $r = 1$ and $p = n+1$ (the numbers of equivalence classes defined by $\mathcal{R}$ and R respectively). The subgroup H is the group of all permutations of the arguments.

THEOREM 3

"Every function f, symmetric in the n_s variables a, can be expressed uniquely as

$$f(a, b) = S(a) g(b)$$

where $S(a)$ is a symmetric function of the variables a only and $g(b)$ is a function of the remaining variables."

It follows from the definitions that every symmetric function is a sum of elementary symmetric functions. It also follows that the product of an elementary function and any other function of an elementary E-algebra is either zero or the elementary function itself.

THEOREM 4

"If Γ_i is an elementary function and $\Gamma_i = S(a)g(b)$ where $S(a)$ is a symmetric function, then $S(a)$ is elementary."

Proof: Suppose not, then it is a sum of elementary symmetric functions

$$S(a) = \sum_j S_{k_j}(a)$$

whence

$$\Gamma_i = g(b) \sum_j S_{k_j}(a)$$

But the function

$$h(a, b) = S_{k_1}(a) g(b)$$

clearly belongs to the elementary E-algebra and

$$\begin{aligned} h(a, b)\Gamma_i(a, b) &\neq 0 \\ &\neq \Gamma_i(a, b) \end{aligned}$$

This contradicts the hypothesis that $\Gamma_i(a,b)$ is elementary.

THEOREM 5

"Every elementary function can be expressed as

$$\Gamma_i = \prod_{k=1}^{r} S_{i_k}(a_k)$$

where $S_{ik}(a_k)$ is an elementary symmetric function of the variables a_k of the kth $\mathcal{R}$-

---

Paper 60-1266, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9-14, 1960. Manuscript submitted June 7, 1960; made available for printing November 23, 1960.

WARREN SEMON is with Sperry Rand Research Center, Sudbury, Mass.

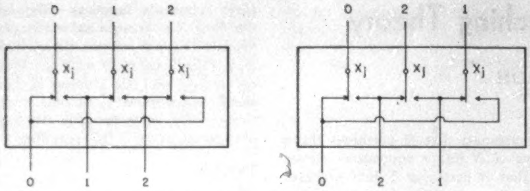


Fig. 1. Subnetwork types

Type A

Type B

equivalence class. Here r is the number of such $\mathcal{R}$-equivalence classes."

Proof: Γ_i is symmetric in the variables $\{a_1\}$. Hence by Theorem 3,

$$\Gamma_i = S(a_1)g_1(a_2, \ldots, a_n)$$

and by Theorem 4, $S(a_1)$ is elementary, and

$$\Gamma_i = (S_{i_1}(a_1)g_1(a_2, \ldots, a_n)$$

Clearly, $g_1(a_2, \ldots, a_n)$ is symmetric in the variables $\{a_2\}$, so that

$$g_1(a_2, \ldots, a_n) = S(a_2)g_2(a_3, \ldots, a_n)$$

Hence

$$\Gamma_i = S(a_2)[S_{i_1}(a_1)g_2(a_3, \ldots, a_n)]$$

and applying Theorem 4 again gives that $S(a_2)$ is elementary. An obvious inductive argument can now be used to complete the proof. It is a corollary that the number p of elementary functions is

$$p = \prod_{i=1}^{r} (k_i + 1)$$

where k_i is the number of indices in the ith $\mathcal{R}$-equivalence class, $i = 1(1)r$.

THEOREM 6

"All the elementary functions of an elementary E-algebra can be realized by a network using

$$\sum_{j=1}^{r} \frac{p_j k_j}{2}$$

transfers, where

$$p_j = \prod_{i=1}^{j} (k_i + 1)."$$

Proof: By induction on r; if $r = 1$ the elementary E-algebra contains the symmetric functions, $p = n+1$, $k = n$ and this is a known result.[1,2] Now, every product of a symmetric function of the k_r variables in the rth equivalence class and an elementary function of the variables in the remaining equivalence classes is an elementary function by Theorem 5. All symmetric functions of the variables in the rth equivalence class can be realized by using $(k_r(k_r+1))/2$ transfers. Each function will be needed p_{r-1} times. Hence the total number of transfers required will be

$$p_{r-1} \cdot \frac{(k_r+1)(k_r)}{2} + \sum_{j=1}^{r=1} \frac{k_j p_j}{2}$$

by the inductive hypothesis. This becomes

$$p_r \frac{k_r}{2} + \sum_{j=1}^{r-1} \frac{k_j p_j}{2} = \sum_{j=1}^{r} \frac{k_j p_j}{2}$$

as asserted.

Since each $p_j \leq p_r$, then

$$\sum_{j=1}^{r} \frac{k_j p_j}{2} \leq p_r \sum_{j=1}^{r} \frac{k_j}{2} \leq \frac{n p_r}{2}$$

and except in trivial cases the inequality is strong.

The two limiting cases of elementary E-algebras are those induced by making $\mathcal{R}$ either the universal relation or the identity relation. The first case produces the symmetric functions and has already been described. In the second case $i\mathcal{R}j$ if and only if $i = j$. Then R becomes the identity relation on Γ, the subsets Γ_i are the configurations themselves, and the induced elementary E-algebra contains all functions. The subgroup H contains the identity alone and the "iterative network" realization is the tree network.

An elementary E-algebra may be modified by superimposing a second equivalence relation on the set $\{\Gamma_i\}$ whose elements are the equivalence classes induced by R. The resulting partitioning of Γ will be said to define an "E-algebra." An example of an E-algebra which is not elementary is produced by superimposing on the "k out of n" relation which defines the symmetric functions, the additional equivalence relation $|k|_2 = 1$. (Here $|\alpha|_m$ is defined to be the least positive residue of α modulo m.) The induced E-algebra contains the alternating symmetric function, its inverse, and the functions 0 and 1. Note that there is no change in $\mathcal{R}$ or H.

Modular Functions

Other examples of nonelementary E-algebras are the algebras of p-modular functions. For these algebras the equivalence relation $\mathcal{R}$ is defined by $i\mathcal{R}j$ if

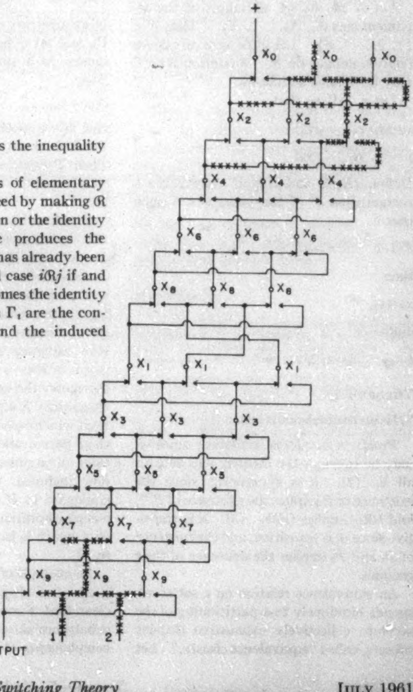


Fig. 2. Iterative network

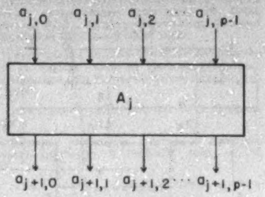


Fig. 3. Typical subnetwork

and only if $i, j \epsilon N$ and $|2^i|_p = |2^j|_p$. The induced relation on Γ may have superimposed on it a relation T, best described in terms of configuration numbers, as kTl if and only if $k, l \epsilon \Gamma$ and $|k|_p = |l|_p$. The elementary p-modular functions are those defined by the sets of configuration numbers

$$\Gamma_a = \left\{ a, a+p, a+2p, \ldots, a+\left[\frac{2^n-a-1}{p}\right]_p \right\}$$

where $a = 0, 1, 2, \ldots, p-1$ (and where the symbol $[\frac{x}{m}]$ means the integral part of x/m).

As an example of a p-modular function, consider the function of ten variables defined by the equations

$$f(x_0, x_1, \ldots, x_9) = \sum_{i=0}^{1023} f_i p_i \qquad (2)$$

where

$$f_i = 1 \text{ for } |i|_3 = 0$$

and

$$f_i = 0 \text{ otherwise} \qquad (3)$$

Equation 2 is, of course, the canonical form for a general 10-variable function as a sum of products, while equation 3 defines the particular function of interest here. This function can be realized by a series chain of subnetworks of the two types shown in Fig. 1. Type A is used for variable x_j if $|2^j|_3 = 1$; type B is used if $|2^j|_3 = 2$. The complete network, employing 48 contacts, is shown in Fig. 2. The crossed lines in the figures are the usual beginning and ending deletions of iterated networks. The function defined by equation 3 is of course the elementary p-modular function obtained by setting $a = 0$ in equation 1, with $p = 3$, $n = 10$. Since the p-modular functions represent a hitherto unremarked set of functions which are realizable by iterative networks, a more detailed discussion is given in what follows.

Suppose that Γ_a is the subset of Γ defining the elementary p-modular function f_{Γ_a}, and $p = 2^k . c$ with c odd. Then in f_{Γ_a} there are k multiplicative variables; if $c = 1$ the remaining n-k variables are vacuous. For example, consider the 4-variable functions defined by $p = 4$. Then $c = 1$, $k = 2$, and

$$\begin{aligned} f_{\Gamma_0} &= x_3'x_2'x_1'x_0' + x_3'x_2x_1'x_0' + x_3x_2'x_1'x_0' + x_3x_2x_1'x_0' \\ &= (x_3'x_2' + x_3'x_2 + x_3x_2' + x_3x_2)x_1'x_0' \\ &= x_1'x_0' \end{aligned}$$

while

$$f_{\Gamma_1} = x_1'x_0,\ f_{\Gamma_2} = x_1x_0',\ f_{\Gamma_3} = x_1x_0$$

Now suppose that $p = 6$, so that $k = 1$, $c = 3$. Then

$$\begin{aligned} f_{\Gamma_0} &= x_3'x_2'x_1'x_0' + x_3'x_2x_1x_0' + x_3x_2x_1'x_0' \\ &= (x_3'x_2'x_1' + x_3'x_2x_1 + x_3x_2x_1')x_0' \end{aligned}$$

In general, the following theorem holds.

THEOREM 7

"If $p = 2^k . c$ with c odd, the elementary p-modular function of n variables f_{Γ_a} can be expressed as

$$f = \dot{x}_{k-1}\dot{x}_{k-2}, \ldots, \dot{x}_1, \dot{x}_0 \{g(x_{n-1}, x_{n-2}, \ldots, x_k)\}$$

where each $\dot{x}_j$, $j = 0(1)k-1$, is either x_j' or x_j according as the coefficient of 2^j is zero or one in the dyadic expression of the integer a; and the function $g(x_k, \ldots, x_{n-1})$ is the c-modular function g_{Γ_b} (as a function of the remaining n-k variables) where $b = [a/2^k]$."

Proof: The integers a and c in dyadic form are:

$$a = a_{n-1}2^{n-1} + a_{n-2}2^{n-2} + \ldots + a_k2^k + a_{k-1}2^{k-1} + \ldots + a_12^1 + a_02^0$$

$$c = c_{n-1}2^{n-1} + c_{n-2}2^{n-2} + \ldots + c_k2^k + c_{k-1}2^{k-1} + \ldots + c_12^1 + c_02^0$$

Since it is assumed that $2^k . c \leq 2^n - 1$

$$2^k \cdot c = c_{n-1}2^{n+k-1} + c_{n-2}2^{n+k-2} + \ldots + c_02^k + 0.2^{k-1} + \ldots + 0.2^1 + 0.2^0$$

All coefficients of powers of 2 above those shown are zero.

The set of dyadic integers representing the integers i for which $f_i = 1$ may be written, by hypothesis, as

$$a, a+1 \cdot 2^k \cdot c, a+2 \cdot 2^k \cdot c, \ldots, a+l \cdot 2^k \cdot c$$

where

$$a + l \cdot 2^k \cdot c \leq 2^n - 1 < a + (l+1) \cdot 2^k \cdot c$$

Each member of this set has the elements $a_0, a_1, \ldots, a_{k-1}$ as coefficients of $2^0, 2^1, \ldots, 2^{k-1}$, respectively; hence the corresponding variables are primed and unprimed in the configuration defined by a in every term. Therefore,

$$f = \dot{x}_0, \dot{x}_1, \ldots, \dot{x}_{k-1}\{g(x_k, \ldots, x_{n-1})\}$$

Note that $c_0 = 1$ so that exactly k variables are multiplicative. Moreover, dividing each i by 2^k gives the set of i for which $g_i = 1$, hence this set is

$$b, b+c, b+2c, \ldots, b+lc$$

where

$$b = \left[\frac{a}{2^k}\right] = a_{n-1}2^{n-k-1} + \ldots + {}_{k+1}2^1 + a_k2^0$$

as asserted.

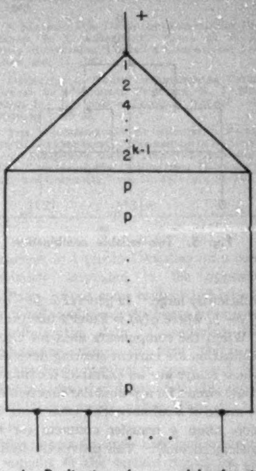


Fig. 4. Realization of a p-modular function

Theorem 7 describes the effect of powers of 2 in the prime factorization of p on an elementary p-modular function f_{Γ_a}. Now consider the case where p is odd. If $2^{n-1} < p$ then an elementary p-modular function consists of a single term. The cases of more interest occur when p is small compared with 2^{n-1}. A p-modular function of n variables, f_M, p odd can always be constructed by iterating subnetworks A_j of the form shown in Fig. 3. The inputs to A_j are p in number, and their values depend on the value modulo p of the dyadic integer corresponding to the configuration of the preceding variables. The subnetwork A_j contains gates dependent on the variable x_j and performs addition modulo p of $|x_j2^j|_p$.

Since it is assumed that p is odd and greater than 1, the powers of 2 will either generate a reduced residue system modulo p exclusive of zero, or part of such a system. Hence there will be at most $p-1$ different subnetworks. More precisely, if 2 belongs to t modulo p, that is, if t is the least positive integer such that $|2^t|_p = 1$, then there will be t different subnetworks, assuming of course that n is

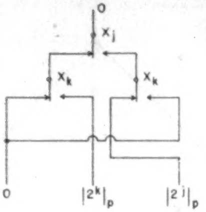


Fig. 5. Two-variable combination

sufficiently large. In general $2 \leq l \leq \phi(p) \leq p-1$, where $\phi(p)$ is Euler's function.

When the components used for circuit realization are current-steering devices of which relays are an example, a combinational circuit for a p-modular function can be made of n subnetworks, each using not more than p transfer contacts (or the analogs thereof). This proves the following theorem.

THEOREM 8

"Any p-modular function of n variables can be realized using not more than np transfers."

This result can be sharpened somewhat by using a tree circuit for the first k variables where $2^{k-1}<p<2^k$. Then, the numbers of transfers are as illustrated in Fig. 4. Hence the function can be realized using not more than

$$N=(2^k-1)+(n-k)p$$

transfers. For an elementary p-modular function the tree may be used for the last k variables as well as the first, so that

$$N=2(2^k-1)+(n-2k)p$$

For the elementary p-modular function shown in Fig. 2, $p=3$, $n=10$. Hence, $k=2$ and

$$\begin{aligned}N&=2(4-1)+(10-4)3\\&=2(3)+6.3=24\end{aligned}$$

transfers, as shown.

Since it is known[5] that the network function defined by $f_i=1$, $i<m$, $f_i=0$ otherwise, can be realized using at most n contacts, the quasi-modular function defined by $f_i=1$, for $i \epsilon \Gamma_a$ and $i<m$, and $f_i=0$ otherwise, can clearly be realized by the series combination of a p-modular function and such a network function. Similarly, step functions[6] can be combined with p-modular functions, and, of course the combinations may be either series or parallel.

For the realization of p-modular functions using voltage-operated devices such as vacuum tubes it is useful to note that if in Fig. 3 all the inputs and outputs are complemented, no physical change need be made in the circuit. This follows from the fact that the inputs and outputs shown in Fig. 3 are related by the equation

$$|a_{j,b+x_j2^j}|_p=a_{j+1,\,|b+x_j2^j|_p}$$

Complementing any input is equivalent to taking the sum of the remaining inputs, thus

$$\begin{aligned}a'_{j,b}=a_{j,0}+a_{j,1}+\ldots+a_{j,b-1}+\\a_{j,b+1}+\ldots+a_{j,p-1}\end{aligned}$$

These two equations may be combined to show that

$$|a'_{j,b+x_j2^j}|_p=a'_{j+1,\,|b+x_j\cdot 2^j|_p}$$

Using this fact, a p-modular function may be realized using not more than $4np$ grids,[4] or analogs thereof.

The bounds given here assume that the variables are treated one at a time and in their natural order. As in other types of iterated networks neither of these restrictions is necessary. It may easily be seen that 2-variable combinations x_j and x_k are chosen so that $|2^j+2^k|_p=0$ have some advantages, since only three outputs result from such a pair instead of the expected four (see Fig. 5). For the function of six variables, with $p=5$, the straightforward method requires 14 transfers = 28 contacts, since $p=5$, $n=6$, $k=3$. However, associating the variables x_0, x_2, x_4 and x_1, x_3, x_5 leads to the circuit shown in Fig. 6, using only 21 contacts, when the dangling branches (shown crossed) are deleted.

Applications

The circuits required for a computer using the residue-number system[8-14] would be largely circuits to realize p-modular functions. Furthermore, it has been demonstrated[12,13] that the carries arising in a fixed or mixed radix adder with composite radices may be described in terms of the carries arising from the same addition carried on modulo the prime factors of the radices. Henderson[14] has shown how an error-correcting code can be constructed using residues, which is preserved under addition and multiplication. He has also discussed the use of p-modular functions in the design of circuits for translation and sign determination in the residue-number system.

Recognition and Identification

Clearly every function which is an element of an E-algebra is invariant to the transformations of a subgroup of the group of argument transformations. Hence any of the methods[15,16] developed for the detection of such invariancies may be used to discover this fact. Henderson has noted in a private communication that decomposition charts[17] may be used for the identification of p-modular functions. The method will probably be applicable to the functions of general E-algebras.

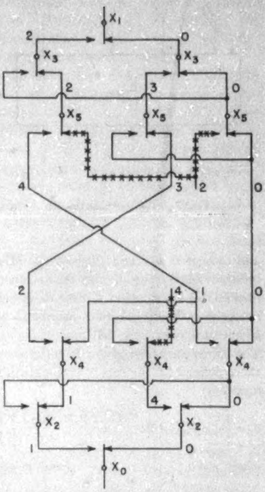


Fig. 6. Modular network

Possession of the requisite symmetries is a necessary but not a sufficient condition for a function to belong to a particular nonelementary E-algebra. A study of the types of equivalence relations which can be superimposed on the elementary E-algebras might lead to the classification of functions according to the E-algebras of which they are elements.

References

1. THE DESIGN OF SWITCHING CIRCUITS (book), W. Keister, A. Ritchie, S. H. Washburn. D. Van Nostrand Company, Inc., Princeton, N. J., 1951.

2. SWITCHING CIRCUITS AND LOGICAL DESIGN (book), S. H. Caldwell. John Wiley & Sons, Inc., New York, N. Y., 1958.

3. LOGICAL DESIGN OF ELECTRIC CIRCUITS (book), R. Higonnet, R. Gréa. McGraw-Hill Book Company, Inc., New York, N. Y., 1958.

4. SYNTHESIS OF ELECTRONIC COMPUTING AND CONTROL CIRCUITS (book). "Annals of the Computation Laboratory of Harvard University," Cambridge, Mass., vol. 27, 1951, pp. 19–20.

5. SYNTHESIS OF SERIES-PARALLEL NETWORK SWITCHING FUNCTIONS, W. L. Semon. *Bell System Technical Journal*, New York, N. Y., vol. 37, July 1958, p. 881.

6. Identification and Optimal Realization of a Class of Switching Functions, M. Cohn. *Report no. BL-21*, "Theory of Switching," Harvard Computation Laboratory, Dec. 1958.

7. Vznik kodu a číselné soustavy zbytkových tříd, M. Valach. *Stroje na spracování informací*, Nakladetelstvi ČSAV, Prague, Czechoslovakia, vol. III, 1955.

8. Operátorové obvody, M. Valach, A. Svoboda. *Ibid.*

9. Rational Numerical System of Residual Classes (in Czech), A. Svoboda. *Ibid.*, vol. V, 1957.

10. The Numerical System of Residual Classes in Mathematical Machines, A. Svoboda. *Proceedings*, Congreso Internacional de Automatica, Madrid, Spain (to be published).

11. The Residue Number System, H. L. Garner. "Proceedings, Western Joint Computer Conference," Institute of Radio Engineers, New York, N. Y., 1959.

12. Advanced Digital Computer Logics, H. H. Aiken, W. L. Semon. *WADC Technical Report*, Wright Air Development Center, Dayton, Ohio, 1956.

13. Advanced Digital Computer Logics, H. H. Aiken, W. L. Semon. *WADC Technical Report no. 59-472*, Wright Air Development Center, 1959

14. Logical Designs for Arithmetic Units, D. S. Henderson. *Ph.D. Thesis*, Harvard University, 1960.

15. A Method for Detecting Functional Invariance, R. L. Ashenhurst. *Report no. BL-2*, "Theory of Switching," Harvard Computation Laboratory, sect. II, 1953.

16. Detection of Group Invariance as Total Symmetry of a Boolean Function, E. J. McCluskey, Jr. *Bell System Technical Journal*, vol. 35, 1956, pp. 1445–53.

17. The Decomposition of Switching Functions, R. L. Ashenhurst. "Proceedings of an International Symposium on the Theory of Switching" (book), "Annals of the Computation Laboratory of Harvard University," vol. 29, 1959, pp. 74–116.

Initiation of Flux Reversal in Magnetic-Amplifier Cores

F. J. FRIEDLAENDER
ASSOCIATE MEMBER AIEE

I. P. LELIAKOV
NONMEMBER AIEE

FLUX REVERSAL in grain-oriented nickel-iron tape cores has been described by several models during recent years.[1,2] Some difference of opinion, however, still exists with regard to the initiation of the process.[1] New evidence is here set forth that nucleation and initial growth of regions or domains of reverse magnetization occur near the tape surface for applied fields sufficiently larger than the coercive force H_c of the material. Fields at least twice the coercive force will be considered. More evidence is still needed to substantiate models for fields only slightly in excess of H_c. Initial growth will be discussed, both for cores previously saturated and for those not saturated when flux reversal was initiated. The latter case, though of particular interest in many applications of "square-loop" magnetic cores, has not been treated adequately so far.

Assumed Model

The rate of change of flux $d\phi/dt$ in a tape core, which has its flux reversed by a step function of magnetomotive force (mmf), is shown in Fig. 1. (The rate of change is determined by noting the voltage across a toroidal winding on the core. The initial voltage spike, which is usually observed, is ignored in this paper since it is not of importance in the processes considered here.) If a model of surface growth of domain walls is used,[2] the initial domain wall geometry takes the form of Fig. 2.

Obviously, for a given value of applied field H, $d\phi/dt$—and hence, the induced voltage per wrap of a given core size—should be independent of tape thickness until domains growing from opposite surfaces of the tape "collide," annihilating the domain walls. The peak voltage E_p (volts per wrap), and hence maximum rate of change of flux, occurs just before merging of adjacent domains is completed; i.e., the peak occurs just before maximum domain wall area is reached. A method permitting experimental corroboration of this fact was recently supplied by Becker.[3] Details of the dynamics of domain wall motion,[4] which have to be considered to explain this result, as well as the experimental evidence,[5] will be presented in later papers.

If such merging occurs before the collision, then both the time to peak t_p and the peak voltage per wrap of tape E_p must be independent of tape thickness for a given value of H. Fig. 3 shows both E_p and the initial voltage per wrap E_i. Fig. 4 displays $1/t_p$ as a function of the field H which is applied by means of a step of current in a toroidal winding on a core previously saturated in the opposite direction. Four cores each of 1/2-mil, 1-mil, 2-mil, and 4-mil thickness were used to obtain the data. Considering the eight-to-one range in tape thickness and the large number of cores used, the indicated spread of data is tolerably small about the average value given in the figure.

Figs. 3 and 4 supply excellent evidence in support of the surface model for the initiation of flux reversal. Also, Fig. 4 clearly indicates that in 1/2-mil tapes, collision of domain walls growing from opposite tape surfaces starts to occur just before the peak voltage is reached, resulting in somewhat larger values of $1/t_p$ than in the case of the thicker cores, for the same values of H.

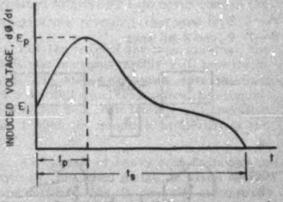


Fig. 1. Rate of change of flux in a previously saturated tape core on application of step of mmf in reverse direction

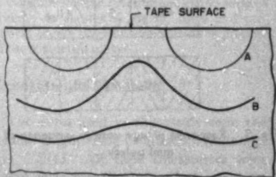


Fig. 2. Approximate initial domain wall geometry in tape core for $H \gg H_c$

A—$0<t<t_p$
B—$t \cong t_p$
C—$t_p<t<t_s$

Paper 60-874, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Atlantic City, N. J., June 19–24, 1960, and re-presented for discussion only at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted April 15, 1960; made available for printing December 7, 1960.

F. J. Friedlaender is with Purdue University, Lafayette, Ind., and I. P. Leliakov is with Aerospace Corporation, El Segundo, Calif.

The work described was made possible through a grant of the National Science Foundation and the Allegheny Ludlum Steel Corporation. Portions of the paper were abstracted from a dissertation submitted by I. P. Leliakov in partial fulfillment of the requirements for the Doctor of Philosophy degree at Purdue University.

The authors would like to thank the Arnold Engineering Company and especially their chief engineer, J. Mitch, for donating the cores used in this work. Professor P. Rautala of Purdue University deserves credit for supplying the metallurgical data, including Fig. 9, through a project carried out by a group of his students. Also, comments by Dr. C. P. Bean, General Electric Company, Schenectady, N. Y., are gratefully acknowledged.

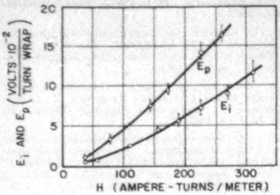


Fig. 3. Peak-induced voltage E_p and the initial-induced voltage E_i, per wrap, as function of applied field H for 1/2-, 1-, 2-, and 4-mil tapes

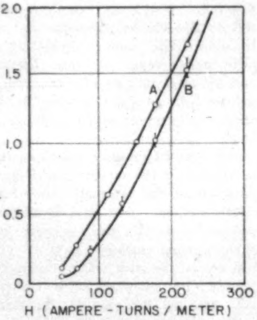


Fig. 4. Reciprocal of time to peak voltage (t_p) as function of applied field H

A—1/2-mil tape
B—1-, 2-, and 4-mil tapes

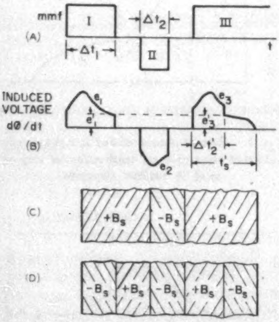


Fig. 5. Response of tape core to sequence of mmf pulses

A—Sequence of mmf pulses applied to core
B—Predicted induced voltages in core (general shape)
C—Approximate domain wall geometry at end of pulse I
D—Approximate domain wall geometry at end of pulse II

Additional Evidence for Model

The model presented here, as can be seen from the foregoing considerations, will give rise to essentially planar domain walls in tapes of 1-, 2-, and 4-mil thicknesses if fields of sufficient amplitude and duration are applied. This leads to a simple domain wall geometry, and readily permits prediction of the relationship between induced voltages for a sequence of mmf pulses applied to a core. Each pulse is long enough to form a simple, planar domain wall geometry, but not long enough to cause collision between planar walls growing from opposite surfaces of the tape, i.e., only partial reversal of flux results.

A core initially saturated in the negative direction (where the usual convention is used that positive currents give rise to positive values of *H* and, hence, positive saturation) has a sequence of current pulses applied to a toroidal winding on the core. These pulses produce the mmf's shown in Fig. 5(A). All mmf's are of equal amplitude, but the second pulse is shorter than the first and of opposite polarity. The third is positive like the first, and of enough duration to drive the core to positive saturation. Even the short pulse II is made long enough to produce essentially a planar wall.

The model discussed here leads to the prediction that the domain wall geometry at the end of pulse I can be approximated as shown in Fig. 5(C), and after II as in 5(D). The predicted general shape for rates of change of flux in the tape (i.e., the induced voltages) are sketched in Fig. 5(B). If flux reversal initiates at the tape surface only, the instantaneous voltages during the first Δt_2 seconds of each of the three pulses should be identical, though e_2 has the opposite polarity to e_1 and e_3.

After Δt_2 seconds of pulse III, however, the planar domain walls reach the region of the tape left saturated positively; see Fig. 5(D). Hence, flux reversal must proceed by further motion of the inner walls left in the tape because of the previous application of pulse I. The induced voltage e_3, therefore, should drop to a value e_3' at this instant, equal to the value e_1' at the end of pulse I.

The oscillograms of Fig. 6 confirm these predictions; (A), (B), and (C) show the induced voltages attributable to pulses I, II, and III, respectively, all to the same scale. In Fig. 7(A), voltages due to pulses I and III are superimposed. In Fig. 7(B), an enlarged portion of Fig. 7(A) is shown for the tail of the pulses,

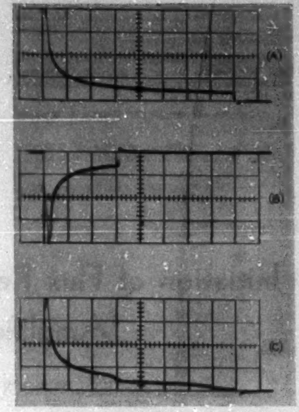


Fig. 6. Oscillograms showing observed induced voltages

A—Pulse I
B—Pulse II
C—Pulse III
Vertical scale: 8.5 volts per division
Horizontal scale: 2.26 microseconds per division

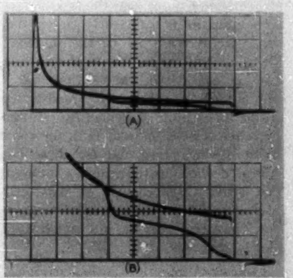


Fig. 7. Oscillograms

A—Voltages for pulses I and III superimposed
B—Enlarged picture (4×) of tail of traces shown in A

clearly indicating that the predictions are quite accurate.

Conclusions to be drawn from this experiment or many similar experiments which may be devised and carried out are enumerated here; for sake of conciseness, details of only one experiment are given.

1. **Flux reversal always initiates near the tape surface for applied fields larger than about twice the coercive force. Whether this statement is true for fields almost equal to the coercive force has not been determined.**

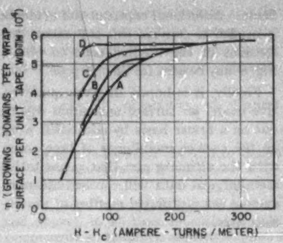


Fig. 8. Number of growing domains as function of $(H-H_c)$ for one wrap of tape, core T8043

A—H applied to saturated core
B, C, D—H applied after prepulse of magnitude given by intercept with A

2. Domain walls in the interior of the tape will not contribute to flux reversal until the region between the walls and the tape surface has been saturated in the direction of the applied field. In other words, all internal walls appear to remain stationary until they are no longer shielded by walls nearer the tape surface.

The range of experimentation over which these results have been confirmed for grain-oriented 50% nickel-iron (Ni-Fe) cores is about $2H_c<H<100H_c$.

Surface Nucleation Sites Contributing to Reversal

As shown previously,[2] the nuclei of reverse magnetization are assumed to be cylinders which intercept the tape surface at a diameter such that, in a cross-sectional view, they appear as shown in Fig. 2(A). The initial growth of the nuclei is given by[2]

$$\beta_r \frac{dr}{dt} = H - H_c \qquad (1)$$

where β_r is a constant depending on spin relaxation and microscopic eddy currents. The r is the average radius of the nuclei in meters, and r_i is the initial average radius for a saturated core.

If there are n growing nuclei per wrap surface per unit width of tape, then the initial rate of change of flux per wrap

$$\left.\frac{d\phi}{dt}\right|_{r=r_i} = E_i = \frac{d}{dt}\,[nw\pi r^2 B_s] \qquad (2)$$

where w is the tape height ("tape width") in meters.

B_s is the saturation flux density in webers/m² (webers per square meter)

$$E_i = 2nw\pi B_s r_i \frac{dr}{dt} \qquad (3)$$

Substituting equation 1 into 3

$$E_i = Kn(H-H_c) \qquad (4)$$

where $K=2w\pi B_s\, r_i/\beta_r$, a constant for a given core which can be determined.[4] Equation 4 can be used to determine n as a function of H, directly from experimental observations of E_i.

In curve A of Fig. 8, n is shown as a function of $(H-H_c)$ as determined from the values of E_i shown in Fig. 3. Since the number of nuclei per wrap of tape (n) is large, n can be approximated by the continuous function

$$n = N[1-e^{-a(H-H_c)}] \qquad (5)$$

This function can be made to fit curve A quite well by means of a suitable choice of a and N.

The value of n for large values of H approaches N. Apparently only N nucleation sites are available from which domains of reverse magnetization can grow, as was suggested previously.[3]

The value of E_i, and hence of n, does not depend on the previous history of applied fields as long as only interior walls exist initially. Different results are obtained if the step of H is preceded by a short "prepulse" H' where $H'>H$. The prepulse must be short enough to obtain essentially the domain wall geometry of Fig. 2(A) at the end of the prepulse. If n is now determined experimentally for different values of H, applied immediately after H', curves B, C, and D of Fig. 8 are obtained. For $H>H'$, curve A is retraced. The magnitude of $(H'-H_c)$ can be determined in each case from the intercept of curve B, C, or D with curve A.

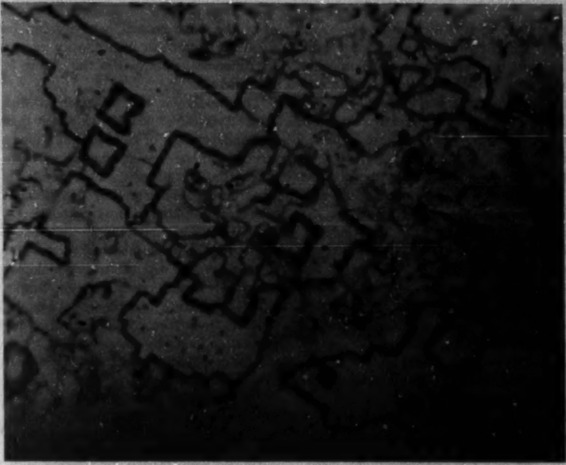


Fig. 9. Partly etched tape surface showing disturbed metal layer, magnification 800×

Metallurgical Considerations

The metallurgical data on the grain-oriented 50% Ni-Fe tape cores used in the experiments revealed that:

1. For the bulk of the tape material, the grain size increases rapidly with tape thickness. As similar responses were obtained with various tape thicknesses, the conclusion is that bulk grain size has no appreciable effect upon the magnetic properties of the core.

2. As a result of heavy cold reduction, a layer of "disturbed metal" is found on the tape surfaces, Fig. 9, and the amount increases with the degree of cold reduction.

3. While X-ray diffraction patterns show good grain orientation of the bulk of the tape material, only qualitative information is available on the orientation of the surface layer. As shown in Fig. 9, some undetermined degree of grain orientation is present in the disturbed metal.

Experimental Methods

The cores used in this study were the Arnold Engineering Company's Deltamax size *T8043*. All tape thicknesses were rolled from the same heat. Cores were made available, in each thickness, with a range of anneal temperature from 1,850 F (degrees Fahrenheit) to 2,050 F in 50 F steps. The best consistency was obtained with cores annealed at 1,950 F, but the variation between those annealed at other temperatures was not great. All cores had been subjected to the usual tests by the manufacturer, and only those meeting minimum specifications for squareness, coercive force, etc., were supplied.

The experimental data presented in the curves of Figs. 3 and 4 were obtained from four cores each of 1/2-mil, 1-mil, 2-mil, and 4-mil thickness, annealed at 1,950 F, and they were driven with programmed sequence of rectangular current pulses through 10-turn toroidal windings. Rates of change of flux were observed by means of a Tektronix, Inc., model *545* oscilloscope connected across an additional winding placed on the core. The programmed current pulses came from a modified Rese Engineering Company's magnetic core tester model *1100*, with current rise and decay times of about 0.09 microsecond, and a constant current output amplitude for the entire duration of the current pulse (within 5% under the worst conditions encountered during the experimentation). The wraps in each core were counted with the help of a microscope, as this information was needed for calculation purposes.

Conclusions

Evidence shows that flux reversal in Ni-Fe tape cores initiates at the surface. While experimentation was carried out with cores made from 50% Ni-Fe tapes only, the same general behavior can be expected from other Ni-Fe tapes, except for differences discussed in the closure of reference 2.

The model for the initiation of flux reversal is developed by showing that it is independent of the internal domain structure of the tape due to the previous magnetization history. This result is important in understanding flux changes which are initiated in unsaturated cores. It now becomes possible to develop a model for the instantaneous process during flux reversal with arbitrary (or almost arbitrary) initial conditions—in simple situations, merely by extending the model presented in reference 2, and in more complex situations, by means of a more careful consideration of the domain wall dynamics.[4]

The results obtained by prepulsing a core with a field of magnitude H' before continuing flux reversal with a smaller field H (i.e. $H<H'$) indicate that domains will continue to grow, once started, even if H is reduced below the normal threshold value for the domain after the initiating pulse is applied. This observation supports the conjecture that the value of starting field required for a given domain depends on its initial radius.[6] The initial pulse apparently has expanded the nuclei enough so that they will continue growing with smaller applied fields. Additional experimental evidence is needed to confirm this statement; extensions of the method used to obtain Fig. 8 may be used for this purpose.

Finally, it is hoped that the quantitative model for surface nucleation will be put on a firmer basis by additional work on the surface metallurgy of tape cores, with the ultimate aim that quantitative metallurgical data will confirm the presence of the postulated number of nucleation sites.

References

1. A Mathematical Model for a Magnetic-Amplifier Reactor Core, H. H. Woodson. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, 1959 (Jan. 1960 section), pp. 872–78. (Additional references quoted in this reference.)

2. Flux Reversal in Magnetic Amplifier Cores, F. J. Friedlaender. *Ibid.*, vol. 75, July 1956, pp. 268–78. (Additional references quoted in this reference.)

3. Domain Boundary Configurations During Magnetization Reversals, J. J. Becker. *Journal of Applied Physics*, New York, N. Y., vol. 30, no. 3, Mar. 1959, pp. 387–90.

4. An Improved Model for Flux Reversal in Ni-Fe Cores, I. P. Leliakov, F. J. Friedlaender. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 80, Mar. 1961, pp. 23–26.

5. Reversible and Slow Speed Irreversible Magnetization Processes in 50% Ni-Fe Alloys, L. L. Ogborn. *Ph.D. Thesis*, Purdue University, Lafayette, Ind., Jan. 1961.

6. The Process of Flux Reversal in Grain-Oriented 50% Nickel-Iron Tape Cores, F. J. Friedlaender. *Magnetic Amplifier Technical Report no. 21*, Carnegie Institute of Technology, Pittsburgh, Pa., Aug. 1955.

Data Gathering and Transmission System for Project EHV

P. A. ABETTI
MEMBER AIEE

D. W. CAVERLY
ASSOCIATE MEMBER AIEE

D. D. DAVIS
MEMBER AIEE

E. A. WHEPLEY
ASSOCIATE MEMBER AIEE

THE General Electric Company extra-high-voltage project (called "Project EHV" for short) was announced in June 1958,[1] after more than a year of preliminary planning. Project EHV involves the design, construction, and operation of a complete a-c transmission system, which will carry power first at 460–500 kv and then at 650–690 kv (750-kv maximum). This is more than twice the maximum voltage presently used by utilities in the United States and Canada and 50% higher than the highest transmission voltage in the world—500 kv in the Soviet Union. Detailed planning, cost estimating, and writing of the specifications for the line and apparatus were essentially completed on May 27, 1959, when construction was begun at the site and in various plants of the co-operating companies. Construction is proceeding and energization at 460–500 kv began in December 1960.

This paper will review briefly the objectives of Project EHV, the reasons for the selection of its right of way as determined by the desired measurements, and finally, the over-all philosophy of system design in relation to the acquisition of data. The data gathering and transmission system, which will obtain about 10 million items of data annually, will be described in detail. Examples will be given of the analog and digital computations which are necessary in order to acquire the planned meteorological, mechanical, and electrical quantities in the desired form. The over-all procedure will be mentioned for the processing of the data and their statistical evaluation through a large digital computer, an International Business Machines Corporation (IBM) *705*.

Previous unpublished papers have presented in detail the over-all technical plan; the preliminary laboratory investigations, such as the insulation design

---

Paper 60-1407, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 12, 1960; made available for printing December 14, 1960.

P. A. Abetti, D. W. Caverly, D. D. Davis, and E. A. Whepley are with the General Electric Company, Pittsfield, Mass.

of the system; the North Station and the three adjacent portal-type towers; the conductors, insulators, and accessories; and the preliminary corona investigations.[2]

The over-all objectives of Project EHV are:

1. To help re-establish the United States as a world leader in the broad areas of EHV power systems.
2. To prove that EHV systems (460 kv and above), in all their interrelated comcomponents and aspects, are feasible, more reliable, and more economical for the transmission of large blocks of power over smaller rights of way than systems presently used.
3. To contribute technical knowledge which will advance the continued expansion of electric power at reduced costs.

Location and System Characteristics

The right of way was selected according to the following specifications:

1. It must be adjacent to an existing 115-kv line of the Western Massachusetts Electric Company, in order to draw power at the two terminal stations. It must be farther than 1,000 feet from existing 115-kv lines in order to avoid radio-noise interference (RI).
2. It must be on relatively level ground, with maximum differences of about 100 feet in elevation. This requirement is dictated by the fact that corona loss, and presumably RI, are dependent upon air density which, in turn, depends upon elevation. For a line of variable altitude, most of the loss and RI would be generated from the sections at the higher altitudes, and it would be quite difficult to analyze and correlate the experimental data.
3. It must be on terrain with a variety of soil conditions (to study its effect on the line performance), and fairly well exposed to lightning, sleet storms, and steady winds (to study aeolian vibrations).

Most of these specifications were met by the site selected—over pastures, swamps, and rocky woods along the Housatonic River, in the towns of Lenox and Lee, Berkshire County, western Massachusetts. The meteorological characteristics of this zone are believed to be representative of many regions of northern United States and southern Canada, and are probably more severe than average United States conditions.

The region has a great variety of weather, with a goodly number of storms of minor intensity, although storms of hurricane force occur occasionally. The temperature varies between −25 F (degrees Fahrenheit) and 95 F. At the location of the line through rural country, contamination other than by pollen should be mild. There is a good deal of fog along the Housatonic River, and it is hoped that critical icing conditions will exist several times during the investigation. The isokeraunic level of the area is 27. While this is rather low for purposes of lightning investigations, is is hoped that about one stroke per mile-year will contact some portion of the line because of the height of the structures which are provided with additional lightning rods. Experiments also will be made to determine the number of strokes per square mile and to obtain correlation with the strokes to the line. The line is equipped with two oscillographs at each of the 18 towers to obtain not only the magnitude but also the waveshape of the stroke currents.

A 4.3-mile 300-foot-wide strip of land was acquired for the line right of way and also approximately 8 acres of land at each end for the terminal stations. Five areas approximately 50×375 feet, uniformly spaced along the right of way, to be used for locating measuring instruments such as RI meters and the analog-to-digital converters, were also obtained. The 300-foot width of the right of way is sufficient for a double-circuit line at 650-kv alternating current; a single-circuit line at 1,000-kv alternating current, or 2×500-kv direct current, which are assumed to be maximum voltages conceivable under present EHV technology. The 4.3-mile length is considerably greater than the length of other experimental EHV lines,[3] which range from 0.03 to 1.4 miles. It was chosen to obtain a better simulation of actual EHV lines, which may have sections from 10 to 200 miles between switching stations. For instance, RI measurements on short lines have shown marked standing waves

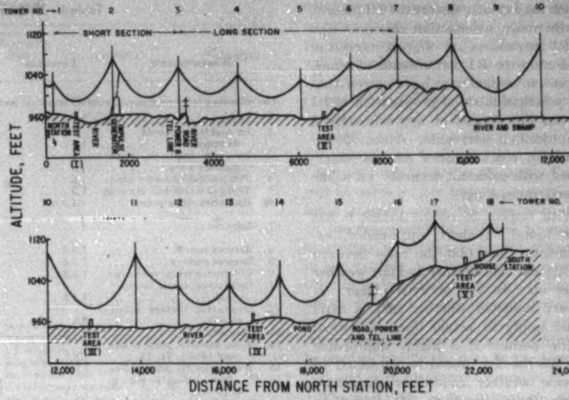


Fig. 1. Profile of Project EHV 650-kv line

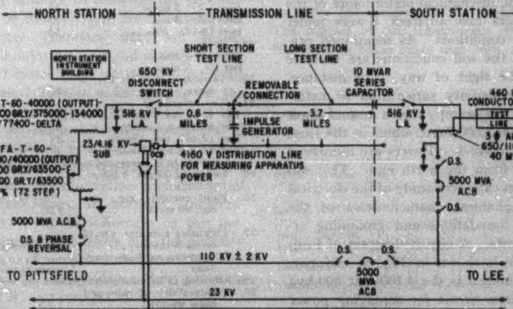

NORTH STATION
TRANSMISSION LINE
SOUTH STATION
NORTH STATION INSTRUMENT BUILDING
650 KV DISCONNECT SWITCH
SHORT SECTION TEST LINE
LONG SECTION TEST LINE
REMOVABLE CONNECTION
10 MVAR SERIES CAPACITOR
460 KV CONDUCTOR AGING TEST LINE
FOA-T-60-40000 (OUTPUT)-650000 GR.Y/375000-134000 GR.Y/77400-DELTA
516 KV L.A.
0.6 MILES
IMPULSE GENERATOR
3.7 MILES
516 KV L.A.
3 ϕ AUTO 650/110 KV 40 MVA
23/4.16 KV SUB
OCB
4160 V DISTRIBUTION LINE FOR MEASURING APPARATUS POWER
OA/FA-T-60-30000/40000 (OUTPUT)-110000 GR.Y/63500-110000 GR.Y/63500 ±30% (72 STEP)
D.S.
5000 MVA A.C.B.
D.S.
5000 MVA A.C.B.
D.S. & PHASE REVERSAL
110 KV ± 2 KV
D.S.
D.S.
5000 MVA ACB
TO PITTSFIELD
TO LEE.
23 KV

Fig. 2. Project EHV circuit diagram

which may greatly distort the RI picture. Furthermore, attenuation characteristics at RI frequencies are of great interest in predicting the RI performance of a line. Recent investigations[4] have shown that the principal mode of propagation of RI has an attenuation of approximately 6 db/ (decibels per) mile. Thus, its attenuation characteristics may be determined with sufficient accuracy on a line several miles long.

With respect to corona losses, it has been found on short, open-ended, experimental lines that the ratio between the highest foul-weather and fair-weather losses may be 100 or even higher. It is very doubtful whether such a ratio would apply to actual lines of normal length, because it is highly improbable that all the line would be subjected to extreme weather conditions simultaneously. It is, therefore, of interest to study the effect of nonuniform weather on the corona, mechanical, and RI performances. In the Berkshires, heavy rain or snow in one location and sunshine in a location a few miles away is quite common. Thus, the Project EHV line, with three complete weather stations located about 2 miles apart, will be very useful in investigating this effect.

Another reason for choosing the 4.3-mile length was to study attenuation characteristics of impulse-generator traveling waves along the line and their penetration into the terminal stations. On a line several miles long, a marked attenuation of the impulse waves may be obtained, and abnormal reflections may be avoided.

The profile of the line, shown in Fig. 1, indicates that there will be a total of 18 towers and 17 spans ranging from 880 to 2,200 feet, with an average span of 1,300 feet. It is desirable to have at least one span over 2,000 feet long to study the mechanical performance of the conductors and adjacent towers under unusual conditions. As mentioned previously, the soil conditions are variable along the right of way. For instance, ground resistivity varies from 10 meter-ohms in the swampy region around tower 11, to 300-1,000 meter-ohms in the loose sand at tower 1. Towers are numbered starting from the north end. This will enable an extensive study of the electrical and mechanical characteristics of the various foundations and grounding arrangements. A circuit diagram of Project EHV is shown in Fig. 2. Of particular interest is the 4,160-volt 300-kva distribution circuit for supplying power to the instruments and the various buildings along the line.

Table I. Project EHV Measurements

Measurements	Locations	Measuring Time	Range	Tolerance
The following to be measured every 20 minutes, and every 2 minutes on demand:				
1. Temp.	1,1,3,5	inst.	−40–+40 C*	2°
2. Ice and frost on cold dummy cond.	1,5	inst.	0–6 lb/ft	3%
3. Ice and frost on warm dummy cond.	1,5	inst.	0–6 lbs/ft.	3%
4. Temp of cold dummy cond.	1,5	inst.	−40–+60 C.	2°
5. Temp of warm dummy cond	1,5	inst.	−20–+80 C.	2°
6. Humidity (dew point)	1,1,3,5	inst.	Dewcell temp. −30 to +50 C	2°
7. Infrared	1	avg.	0–2 gm–cal/cm²/min.	3%
8. Aerosol count	1,3,5	inst.	10^5 to 10^7 per ft³	5%
9. Aerosol density	1,3,5	inst.	10^{-9} to 10^{-4} gm/l	10%
10. Conductor temp, cold cond.	1,3,5	inst.	—40–+60 C.	2°
11. Conductor temp, warm cond.	1,3,5	inst.	−20–+80 C.	2°
12. Conductor surface conditions†	1,5	inst.		
13. Pressure	1,3,5	inst.	26.5–30.5 inches	3%
14. Wind velocity, perp. to line near twrs 1, 11, 16	1,3,5	peak	0–75 mph	3%
15. Wind velocity, perp. to line	1,3,5	avg.	0–50 mph	3%
16, 17. Wind velocity, par. to line	1,3,5	peak avg.	0–75 mph 0–50 mph	3%
18. Rainfall and snowfall‡	1,3,5	inst.	0–6 inches	3%
19. Start of rain and snow	1,3,5	time on	0.002 inch	within 1 min.
20. Wind contacts in bundle	1	count		
21, 22. Line voltage	1,5	peak avg.	750 kv, LL	1%
23. Corona loss§	1	avg.	200, 2,000 kw	5%
23A. Porcelain loss, alternate measurement§	1	avg.	50, 500 watts	5%
24. Reactive kva	1	avg.	0–20 kvar	1%
24A. Porcelain, reactive kva, alternate measurement	1	avg.	0–500 var	1%
25. Radio noise:§				
1.04 mc, under heated phase	1,3,5	avg.	44–70 db	3%
1.04 mc, 100 feet out	1,3,5,4	avg.	24–50 db	
1.04 mc, 150 feet out	1,3,5	avg.	17–43 db	
1.04 mc, 100 feet out and 1/4 wave axially	1,3,5	avg.	24–50 db	
1.04 mc, 100 feet out from unheated phase	3	avg.	24–50 db	
1.04 mc conducted	1,3,5	avg.	78–104 db	
1.04 mc conducted, aging area: Single cond. 2 cond. 4 cond.	5 one of three will be made	avg.	66–92 db	
1.04 mc, background	1	avg.	10–70 db	
1.04 mc, 115-kv line	1	avg.	35–61 db	
150 kv, 100 feet out	1,3,5	avg.	38–64 db	
26. Corona pulses	1	count	500 to 1 threshold	5%
27. Electric field	1,3,5	cont. and avg. of log	50 volts to 15 kv/m	20%; 3% of log
28. Aeolian vibration, outside phases: 8 meas, ground wire, 1 meas, twrs 1 and 11	1,3	inst.	10 *g*; 100 cps	3%
29. Axial tension, outside phases each cond: 8 meas, twrs 1 and 18	1,5	peak and avg.	17–25,000 lbs	3%
30. Torque, center phase, 1 cond, twrs 1 and 11	1,3	peak and avg.	1,500 lb in.	3%
31. Sag:				
twrs 1-2: 3 meas	1	both peaks	9°	3%
twrs 2-3: 3 meas	1			
twrs 18-17: 3 meas	5			
32. Swing:				
twr 11: 3 meas	3	peak	35°	3%
twr 16: 3 meas	5			
33. Vertical load: twr 11: 3 meas	3	peak	20,000	3%
34. Twr. stress:				
Twr. 4, Al, tan, 4 meas	2	peak		3%
Twr. 8, st, tan, 4 meas	2			
Twr. 12, con, tan, 4 meas	3			
Twr. 14, wd, tan, 4 meas	4			
The following to be measured as they occur but not processed automatically:				
35. Switching surges, 3 meas	1,5	waveshape	1,800 kv	5%
36. Fault currents, etc., mag osc, 20 meas	1,5	waveshape	160 amps on high side 5,300 amps on low side	5%
37. Lightning currents, osc (2 on each tower)	every tower	waveshape	70 ka	10%
37A. Lightning currents, mag links	every tower	peak amplitude	80 ka	5%
The following to be measured regularly but not processed automatically:				
38. Conductivity of rain and snow weekly‖	1,3,5	avg.	10,000 ohms/cm³	5%
39. Aerosols, high-volume air sampler‖	1	avg.	type and weight	

Table I. (continued)

Measurements	Locations	Measuring Time	Range	Tolerance
40. Small ion meter	1	inst.		
41. Conductor contamination evert 20 min¶	1	cont.		
42. Porcelain contamination every 20 min¶	1	cont.		
43. Aerosols, dust spot every 20 min	1	avg.		
44. TV interference, weekly	1	inst.		20%

* Degrees centigrade.

† Photographically.

‡ 0.005 resolution.

§ Two scales.

‖ Manually.

¶ Closed-circuit TV.

Abbreviations:

Al = aluminum
amps = amperes
avg = average
cir = circuit
con = concrete
cond = conductor
cont = continuous
cps = cycles per second
dia = diameter
gm-cal = gram-calories
gm/l = grams/liter
inst = instantaneous
ka = kiloamperes
kv/m = kilovolts per meter

lb-in = pound-inches
lbs/ft = pounds per foot
LL = line to line
mag = magnetic
mc = megacycles
meas = measurements
mph = miles per hour
osc = oscillograph
par = parallel
perp = perpendicular
st = steel
temp = temperature
twrs = towers
wd = laminated wood

Philosophy and Scope of Measurements

One of the most important goals of Project EHV is to obtain experimental data on an EHV line in actual operation. Advanced measurement systems in connection with a modern digital computer and techniques of statistical analysis will permit attainment of this goal with an unprecedented quantity and accuracy of data. Not only the study of the broad practical aspects will be undertaken, but physical causes for observed effects will be interpreted in as much detail as possible. This should eliminate some of the broad scatter of information from other investigations. For instance, weather has often been classified as fair or foul; we will subdivide foul weather into rain, fog, snow, and frost. Project EHV will try to establish correlation between laboratory and field results, particularly for RIV (RI voltage) measurements, and to establish all the reasons for 2 to 1 lower performance of conductors installed on lines compared with the same conductors tested in the laboratory. In the past, data on most experimental lines frequently have been obtained at irregular times either through observation or recording on charts. Subsequent analysis of the data becomes very laborious because it must first be converted into numerical form.

In Project EHV all data will be measured, transmitted, and processed automatically. Since many conditions beyond the observers' control influence the data, continuous long-duration measurements are required to obtain answers which have statistical significance. For instance, RI may vary in the ratio of 30 to 1 as a result of weather, time of year, or day, etc. A statistical distribution of RI versus time is required, not just the average fair-weather or foul-weather values. With such a statistical distribution, it will be possible to make meaningful predictions. For instance, under actual conditions, RI will not exceed a given value for 95% of the total time. Because of the stochastic nature of most experimental data, it will be necessary to accumulate a quantity of measurements in order to compare the performance of different line configurations for instance. Therefore, data will be acquired automatically and simultaneously on all instruments at definite intervals of time. This is a broadening of the concept first used in the 500-kv test station of Rheinau.[5] All quantities of theoretical or practical interest will be measured for the determination of the electrical and mechanical performance of an EHV line. In addition, measurements of outside influences on this performance will be taken, e.g., the quantity and size of aerosols (small particles floating in the air) because they probably affect RI and corona loss.

A list of all the measurements to be made on Project EHV, their location, measurement time, range, and specified accuracy is given in Table I. A total of 267 measurements of 47 different types will be made, subdivided as follows: 133 (13 types) will be electrical, such as corona loss and RI which will be measured in 23 locations; 71 (11 types) will be mechanical, such as aeolian vibrations, conductor, and tower stresses; and 63 (23 types) will be meteorological. They will be made at five test stations spaced fairly equally along the line (Fig. 1). Of the measurements, 158 will be made at regular intervals: every 20 minutes in fair weather and 2 minutes automatically in bad weather when preset limits are exceeded, or at any time on demand, for a total of about 10 million yearly. An automatic digital system (Fig. 3) will be used to handle the data at the lowest cost and with maximum accuracy.

The first link in the automatic data acquisition system is the transducer, which converts the quantity being measured into an electrical quantity such as voltage, resistance, or switch position. Whenever the output of the transducers is resistive, it is converted to a d-c voltage; for example, accelerometers, whose output is a change of resistance, and barometers where bellows actuate a potentiometer. Typical transducers with output in the form of a d-c voltage are the pyrheliometer for infrared measurement and radio-noise meters. In most cases the transducers are accessible (as for the 23 RI meters) and are connected to the data acquisition system by a system of shielded wires. In some cases the transducers are located directly on the conductors of the EHV line and their output (as for some strain gages) is brought to ground by special radio telemetering equipment. This equipment is designed for operation from batteries which will be recharged every 30 days. It will work in all the severe environmental conditions normally encountered in the Berkshires. The output of the transducers changes very fast in some cases and very slowly

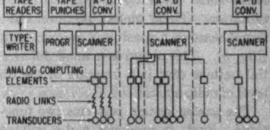


Fig. 3. Simplified block diagram of measurement and data acquisition system

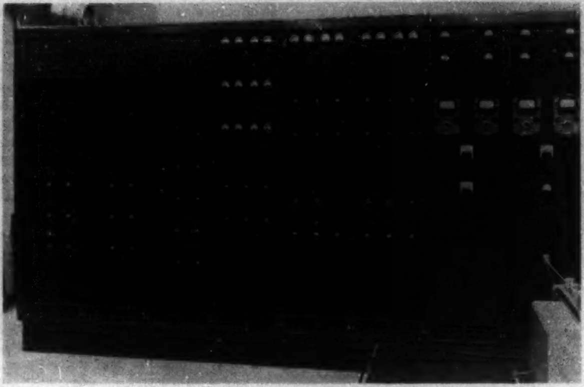


Fig. 4. Analog computing devices for station I

in other cases. The outputs which change quickly could not be acquired and processed economically, and therefore they are processed in analog computers to obtain maximum value, average value, and in some cases, both. For instance, maximums are taken of conductor sag and swing and tower stress; averages are taken of RI and corona loss; both are taken for wind velocity and conductor tension.

Examples will now be given of the methods and circuits employed for performing specific measurements such as radio noise, wind, corona pulses, aeolian vibrations, and corona loss. The analog computing equipment which has been designed for Project EHV will be described in detail for these five particular measurements. The complexity and extent of these analog devices may be seen in Fig. 4, which shows all the analog computing devices which pertain to station I only.

Radio-Noise Measurement

Radio interference is one of the most important measurements on Project EHV. Twenty-three radio-noise meters are located along the line to measure the radio noise either by conduction or by radiation. These meters produce a d-c signal based on a quasi-peak measurement (Fig. 5). This signal is a small differential voltage that is a function of the db level of radio noise.

The analog system performs two functions. First, it produces a d-c voltage representing the average value of the db signal over either a 2-minute or a 20-minute time period. Second, by monitoring the analog output, it changes range and provides an indication of the change.

To perform the first function, the signal from the radio-noise meter must be amplified by a differential amplifier, pass through a compensating network, and be integrated for a fixed period of time. The second function requires the use of a range changer monitoring the output of the integrator. Although this amplifier monitors continually, it does not alter the range except at a specific programmed time. The amplifier changes the attenuation of the radio-noise signal and provides a d-c voltage to the scanner indicating the range.

The differential amplifier used in this measurement is a chopper amplifier with a high rejection of common d-c signals. It is used as a gain amplifier for low-level signals and provides a single-ended output with a low output impedance. This amplifier has high accuracy, low ripple, rapid response, and relatively insignificant drift to the measured signal. The compensating network is a biasing voltage that makes the measured signal a linear function passing through zero. The integrator is a conventional circuit using a high-gain-chopper stabilized amplifier with capacitor feedback. The output will furnish the time integral of the input voltage. The average value of the input voltage will then be represented by the amplifier output divided by the integrating period. This integrator has a high degree of accuracy and will hold the output undisturbed for long intervals at which the input voltage is zero. It has a high input impedance, low output impedance, and very little drift since the operational amplifier is chopper stabilized.

A number of auxiliary features are incorporated as follows:

1. A bistable flip-flop circuit is used for resetting the integrator. The signals operating this circuit are programmed pulses from the scanner which thus control the time of integration.
2. An integrator gain-changing circuit allows the output signal to be independent of the two time cycles. An AND circuit is used to change scales only during the resetting of the integrator.
3. A limiting circuit is employed to keep the amplifier from going into saturation under abnormal conditions. This allows a shorter time for resetting during operations and prevents inaccuracies in the following measurement period.
4. A drift adjustment is provided for normal servicing.

The range changer is an amplifier operating high-speed relays. If the input voltage to this amplifier is outside a predetermined range, the relays, operating as a range decision circuit, control the transducer's level for the next measurement. No switching is done during an operating mode. A bistable flip-flop

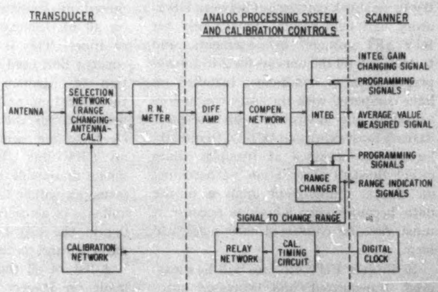


Fig. 5. Radio-noise measurement

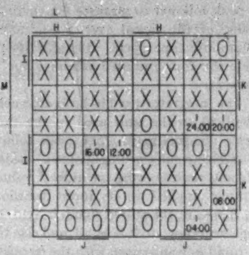


Fig. 6. Six-variable Veitch diagram for 4-hour determination

circuit programmed by the scanner controls the switching function.

Radio-Noise-Meter Calibration

Since the radio-noise meters can get out of alignment for various reasons, e.g., the weakening of bias batteries, they must be supplied with some standard signal to provide a periodic calibration. This calibration signal will replace the normal signal received at the antenna.

Referring to Fig. 5, the requirements for the calibration timing circuit are that: (1) it be turned on at selected times during a 24-hour period; (2) it remain on for 2 or 20 minutes; and (3) that it permit the calibration equipment to be turned on in synchronization with the integrators.

In Fig. 5, a relay network is shown following the calibration timing circuit. The purpose of this network is to assemble in one package the functions: (1) selection of the 2- or 20-minute output of the timing circuit, (2) the final co-ordination of the calibration timing circuit with the integrators, and (3) the selection of a range-changing signal (+48 volts direct current) or a calibration timing signal (−48 volts direct current). The digital clock was selected to provide the time base for the timing circuit. This was the natural selection since the integrators run from this clock and its use satisfied the need for some external means of synchronization.

The radio-noise meters are calibrated every 4 hours when the system is in operation. When the 20-minute time cycle of operation is in effect, the first 20-minute period after the start of the hour is used. The problem is somewhat complicated when the 2-minute cycle of operation is in effect, in that the first 2-minute period after the start of the hour cannot be used. The periods finally chosen were 0008–0010, 0408–0410, 0808–0810, 1208–1210, 1608–1610, and 2008–2010.

The output of the digital clock is binary-coded decimal. The output lines of the clock are numbered and the time may be indicated by simply writing these numbers. The problem of design was simplified by dividing the problem into two parts and combining the two solutions. The problem of turning on the calibration equipment every 4 hours can be solved first since it is common to both the 2- and 20-minute cycle. The method of design is to write the time in local form, transfer this logical form to a truth table written for the maximum number of variables, and then use a Veitch diagram to arrive at the correct solution.

The hourly output of the clock can be defined by the following lines:

$$\text{Output}_{(\text{hours})} = f_b\,(\text{lines } H, I, J, K, L, M) \quad (1)$$

This means that any one of the 24 hours can be defined by the presence of one or more of the variables in equation 1. If the binary number 1 is assigned to the lines when a potential of +6 volts is present and binary number 0 is assigned when no voltage is present, the time of the day can be written as follows.

$$\begin{aligned} 0400 &= J = 001000\\ 0800 &= K = 000100\\ 1200 &= I, L = 010010\\ 1600 &= I, J, L = 011010\\ 2000 &= M = 000001\\ 2400 &= J, M = 001001 \end{aligned} \quad (2)$$

The binary truth table for six variables was prepared along with the logical form of each of the 24 hours. Since six variables produce a 64-element truth table, there are 58 combinations left after the desired 6 calibration hours are noted. Eighteen of the 58 combinations are the remaining 18 hours of the day. However, these hours are not desired as calibrating periods, thus the 18 possibilities form the so-called "forbidden combinations." The remaining 40 combinations comprise the "redundancies," meaning that they can be used for simplification of the output functions.

The truth table is now transferred to a 6-variable Veitch diagram in which the numbers in equation 1 are equated to letters. This is shown in Fig. 6. The desired 4-hour calibration periods are marked with their hours. The circles indicate the 18 forbidden combinations and the X's are the redundancies.

Using the combination rules of Boolean algebra, the output for the 4-hour calibration period is:

$$\begin{aligned} \text{Output}_{(4\text{-hour})} &= \bar{H}\cdot\bar{I}\cdot\bar{L}+\bar{H}\cdot I\cdot L\\ &= \bar{H}(I\cdot L+\bar{I}\cdot\bar{L}) \end{aligned} \quad (3)$$

where the bar over the letter indicates the NOT function; the dots show that the relay contacts indicated by the letters should be connected in series; and the +'s indicate that the relay contacts should be connected in parallel.

If a circuit were to be fabricated on the basis of equation 3, the calibration period would last an entire hour—an undesirable condition. Now, a solution of a circuit must be found which would be active for only a 2- or 20-minute period. When this circuit is combined with the one just solved, the desired calibration timing circuit will fulfill the first two requirements. The method of solution is very similar to the one used for the 4-hour determination and will not be shown here.

$$\text{Output}_{(20\text{-min})} = \bar{F}\cdot\bar{G} \quad (4)$$

$$\text{Output}_{(2\text{-min})} = D\cdot\bar{E}\cdot\bar{F}\cdot\bar{G} \quad (5)$$

By combining equations 3 and 4, the timing circuit will be active for a period of 20 minutes every 4 hours as indicated by the subscript. The output is:

$$\text{Output}_{(4\ 20\text{-min})} = \bar{H}\cdot(I\cdot L+\bar{I}\cdot\bar{L})+\bar{F}\cdot\bar{G} \quad (6)$$

The combination of equations 3 and 5 will provide an output that is active for 2 minutes every 4 hours as indicated by the subscript. The output is:

$$\text{Output}_{(4\ 2\text{-min})} = \bar{H}\cdot(I\cdot L+\bar{I}\cdot\bar{L})+ D\cdot\bar{E}\cdot\bar{F}\cdot\bar{G} \quad (7)$$

Relays with C-type contacts will implement the Boolean expressions. The letters in equations 6 and 7 are output lines from the clock; therefore the coil of the relay would be energized if the line is active. The normally closed contact of the relay would be designated with the NOT notation ($\bar{A}$, $\bar{B}$, $\bar{C} \ldots \bar{L}$) and the open contact with the normal notation (A, B, $C \ldots L$).

Fig. 7 is derived from equations 6 and 7. The coil of each relay is energized from the output line of the digital clock which corresponds to the letter indicated

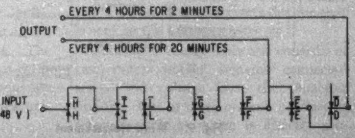


Fig. 7. Calibration timing circuit

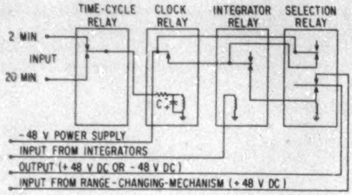

TIME-CYCLE RELAY
CLOCK RELAY
INTEGRATOR RELAY
SELECTION RELAY
2 MIN
INPUT
20 MIN
C
-48 V POWER SUPPLY
INPUT FROM INTEGRATORS
OUTPUT (+48 V DC OR -48 V DC)
INPUT FROM RANGE-CHANGING-MECHANISM (+48 V DC)

Fig. 8. Calibration relay circuit

on the contacts. The contacts are shown in their de-energized position. There is a simultaneous output of the 4-hour 2-minute function and 4-hour 20-minute function since relays *F* and *G* are common to both functions. The output of the timing circuit is the input to the time-cycle relay. This relay selects which time (4–2 or 4–20) function is to be used and is controlled from external equipment. The clock relay consists of seven *A* contacts wired in parallel on the input side. The seven outputs feed the seven integrator relays necessary for the seven radio-noise meters. The integrator relays are energized each time the information integrators operate each 2- or 20-minute cycle.

It is possible that the integrators could overlap the active period of the timing circuit. However, the calibration equipment must be feeding the radio-noise meter for the entire time the integrators operate to take full advantage of the integration period. If the timing circuit is active but the integrator is inactive, no output is desired. If both are active, an output to the calibration equipment is desired. If the timing clock is inactive but the integrator is still active, the output is still "up." Thus, the arrangement shown in Fig. 8 for the clock, integrator, and selection relay top contact. During the periods that no calibration is desired, the timing-circuit relays are continually switching and causing "races." The races cause the clock relay to operate for very short periods. The simplest solution was to place an *RC* (resistance capacitance) circuit ahead of the clock relay.

Wind Measurement

To correlate the effects of the wind with the physical movement of the lines and towers, average and peak values of wind are measured both perpendicular and parallel to the line. An anemometer located at the height of the line provides two signals: (1) a d-c voltage proportional to the velocity of the wind; and (2) the output of a synchronous generator which, when used with a synchronous motor, provides an indication of wind direction; see Fig. 9.

The analog system combines these two signals in the wind direction discriminator. The d-c signal, after filtering, excites a sine-cosine potentiometer, while the synchro signal, through a synchro motor, orients the potentiometer. The two outputs of the sine-cosine potentiometer represent the quadrature components of the velocity of the wind. Using identical circuits, each component signal goes to a high-input impedance amplifier providing an output signal of one polarity. This unit is called the absolute value amplifier. The signal from this amplifier passes through an integrator to calculate the average value, and through a peak follower to measure the maximum value of the signal over a 2-minute or 20-minute time cycle. These units provide the four required readings for the scanner.

The integrator used in both channels is identical to the unit used in the radio-noise measurement. The peak follower is an electronic circuit capable of measuring and remembering the highest value of voltage applied to its input terminals. In this unit, the stored signal is held with high accuracy, drift is small, and rise time is rapid in comparison with the frequency and amplitude of the phenomenon being recorded.

The peak follower consists of two chopper stabilized amplifiers and the associated diodes. The first is used as a gain amplifier and the second as a peak storage unit. Feedback is employed to compensate for the diode voltage drop within the circuit. A limiting circuit is used to keep the amplifiers from going into saturation. This peak follower must be reset after each time cycle. Resetting is accomplished with a circuit similar to that used in the integrator, and programming is again controlled by the scanner. The peak follower has several additional features making it very versatile. It can, with a few minor changes, be used to measure and remember the lowest value of the input voltage (minimum peak measurement). This application is employed elsewhere in the analog system. In addition, this device can have either a gain of 40 or of 1.

Corona-Pulse Measurement

One of the measurements associated with the effects of corona is the counting of high-frequency negative pulses according to their magnitudes. The pulses generated at the high-voltage line are conducted to the analog processing equip-

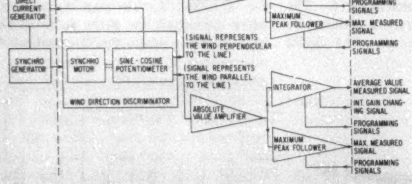

TRANSDUCER
ANEMOMETER
DIRECT CURRENT GENERATOR
SYNCHRO GENERATOR
ANALOG PROCESSING SYSTEM
SYNCHRO MOTOR
SINE-COSINE POTENTIOMETER
WIND DIRECTION DISCRIMINATOR
(SIGNAL REPRESENTS THE WIND PERPENDICULAR TO THE LINE)
(SIGNAL REPRESENTS THE WIND PARALLEL TO THE LINE)
ABSOLUTE VALUE AMPLIFIER
INTEGRATOR
MAXIMUM PEAK FOLLOWER
ABSOLUTE VALUE AMPLIFIER
INTEGRATOR
MAXIMUM PEAK FOLLOWER
SCANNER
AVERAGE VALUE MEASURED SIGNAL
INT. GAIN CHANGING SIGNAL
PROGRAMMING SIGNALS
MAX. MEASURED SIGNAL
PROGRAMMING SIGNALS
AVERAGE VALUE MEASURED SIGNAL
INT. GAIN CHANGING SIGNAL
PROGRAMMING SIGNALS
MAX. MEASURED SIGNAL
PROGRAMMING SIGNALS

Fig. 9. Wind measurement

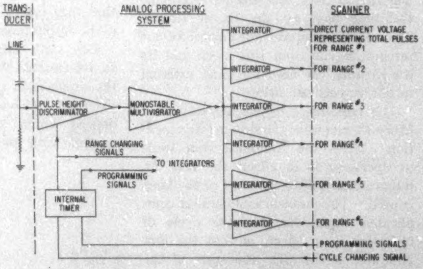

TRANSDUCER
LINE
ANALOG PROCESSING SYSTEM
PULSE HEIGHT DISCRIMINATOR
MONOSTABLE MULTIVIBRATOR
RANGE CHANGING SIGNALS
TO INTEGRATORS
PROGRAMMING SIGNALS
INTERNAL TIMER
INTEGRATOR
INTEGRATOR
INTEGRATOR
INTEGRATOR
INTEGRATOR
INTEGRATOR
SCANNER
DIRECT CURRENT VOLTAGE REPRESENTING TOTAL PULSES FOR RANGE #1
FOR RANGE #2
FOR RANGE #3
FOR RANGE #4
FOR RANGE #5
FOR RANGE #6
PROGRAMMING SIGNALS
CYCLE CHANGING SIGNAL

Fig. 10. Corona-pulse measurement

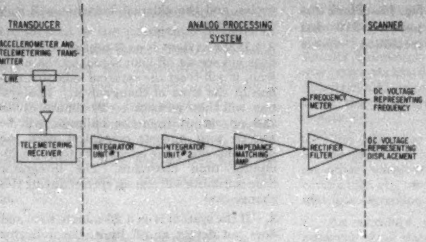


Fig. 11. Aeolian vibration measurement

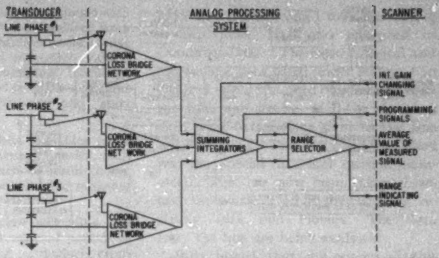


Fig. 12. Corona-loss measurement

ment by a capacitor-resistor divider. The processing equipment counts the number of pulses within specified voltage limits. These limits are 0–50, 50–100, 100–200, 200–300, 300–400, 400–500. This produces six d-c voltage outputs which correspond to the total number of pulses counted in each group. Each magnitude range has its own rate of repetition. Thus, the counting period for each group is made inversely proportional to the maximum average repetition rate, making the maximum output level the same for each counter.

The process equipment consists of a pulse height discriminator (magnitude range selector), a monostable-multivibrator, six integrators, and an internal timing device (Fig. 10). The pulse height discriminator controls the range selection for the counting of different magnitude pulses. The discriminator produces an output pulse which triggers a monostable-multivibrator for each pulse detected in the selected range. The multivibrator's output is passed through a "clipper" circuit and then directed to an integrator. An internal timing circuit co-ordinates the selection of an integrator with the corresponding magnitude range of the pulse height discriminator. The timer also controls the resetting functions of the equipment after each period of operation. The program signals from the scanner control the internal timing circuit to synchronize the operation of the measurement with the scanner.

Since the corona pulse system's entire operation for one measurement period is 2 minutes, it operates continuously when the scanner is on a 2-minute cycle and only during the last 2 minutes of a 20-minute cycle.

The pulse height discriminator, using two Schmitt triggers, sets the upper and lower limits of the magnitude range. The output of these circuits, working into an anticoincidence circuit, produces an output signal dependent on the input signal and the magnitude range setting. The integrators are basically the same as those used in the radio-noise measurement. They do not require extra circuits for programming and integrator gain changing since the internal timer controls the time function. The internal timer consists of two mechanical timers, the first delays the measurement for 18 minutes on a 20-minute cycle and the second controls the time cycle of the operation.

Aeolian Vibration Measurement

Aeolian vibration is the low-frequency harmonic vibration of conductors set in motion by the wind. Measurements of the displacement of the conductor from its normal position and the frequency of the motion are taken once during each measurement cycle; see Fig. 11.

An accelerometer is used as the transducer to measure this vibration. Since it is attached to the high-voltage line, the output signal must be telemetered to the analog processing system. This signal is a function of the acceleration of the line and is integrated twice by the analog equipment to provide a voltage proportional to displacement. An impedance-matching amplifier raises the power level of the signal so that it can feed two devices: (1) a frequency meter whose output provides a d-c signal corresponding to the frequency of the vibration; (2) a rectifier filter and load system which provides a d-c signal indicating the displacement of the conductor. These two output signals are sent to the scanner.

The processing equipment is designed to be used with either an accelerometer or a strain gage. In the latter case, the signal by-passes the integrator units. The two integrator units consist of an a-c amplifier and an integrator network. The frequency-response characteristic of these units is designed so that low-frequency disturbances do not affect the measurement.

Corona-Loss Measurement

Measurement of the corona loss has been described by Foley and Olsen.[3] Self-balancing bridge networks measure the loss for each phase. Since the average value of the total loss for an operating cycle is required, an electrical signal representing the output loss for each network is fed to the summing integrators. The voltage from the integrating units represents the average value of the 3-phase loss. This voltage signal and a range indicating signal are sent to the scanner; Fig. 12.

Each bridge network has two inputs: (1) a voltage proportional and in phase with the line-to-ground voltage; (2) a voltage that is proportional and in phase with the current in the high-voltage line. The first voltage is obtained from a precision capacitor divider and the second from a current transformer and burden resistor located with the line. The latter signal must be telemetered to the ground equipment. The bridge networks produce two outputs that are related to the impedance of the measurement. These outputs are calibrated in watts and reactive volt-amperes (vars) since the line-to-ground voltage is known.

The integrating units consist of three summing integrators and a range selector. Each integrator measures the same incoming signals. Since the corona-loss measurement covers a wide range, each integrator has a different range. On an average input signal the first integrator would read low, the second would read near full scale, and the third would be saturated. The range selector monitors the three outputs and selects the one that is sent to the scanner.

The basic bridge network consists of two potentiometer systems connected across the voltage from the capacitor divider. One of these systems produces a reference voltage which is directly pro-

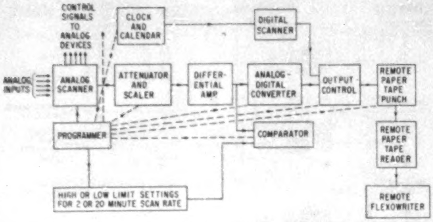

CONTROL SIGNALS TO ANALOG DEVICES
ANALOG INPUTS
ANALOG SCANNER
CLOCK AND CALENDAR
DIGITAL SCANNER
ATTENUATOR AND SCALER
DIFFER-ENTIAL AMP
ANALOG-DIGITAL CONVERTER
OUTPUT CONTROL
REMOTE PAPER TAPE PUNCH
PROGRAMMER
COMPARATOR
REMOTE PAPER TAPE READER
HIGH OR LOW LIMIT SETTINGS FOR 2 OR 20 MINUTE SCAN RATE
REMOTE FLEXOWRITER

Fig. 13. Block diagram of 310 data acquisition system

portional to the line-to-ground voltage. The other system produces a reference voltage which is 90 degrees out of phase with the line-to-ground voltage. The voltage that is telemetered from the current transformer is complex and consists of real and quadrature components relative to the line-to-ground voltage. When the bridge network is balanced, portions of the voltages across the potentiometers are chosen to construct a voltage exactly equal in magnitude but opposite in phase to the voltage from the current. The position of these potentiometers indicates the watts loss and vars in the phase. The summing integrators must change gain for the two time cycles. Program signals from the scanner are used by the integrators and the range selector for all time functions.

Data Acquisition System

For all the 158 measurements to be processed automatically, analog outputs are available at stations I, III, and V. The General Electric Company *310* data acquisition systems will be used to compile this information in digital form on paper tape and on typed log sheets at station I.

Fig. 13 shows a block diagram of one of the three data acquisition systems. The analog scanner can be set to scan once every 20 minutes, once every 2 minutes, or to make a selection automatically. It will scan through 100 differential inputs at 3 per second. Since the input voltages will vary from 10 mv (millivolts) to 100 volts, an attenuator with steps of 1, 10, 100, and 1,000 is programmed on the front-panel pin board and goes to the correct setting as each input is scanned. In order to make the output read in engineering units, a second attenuator called a scaler is also provided. It can be pinned from 10.0 to 99.9. In order to reduce common mode noise and to handle differential outputs, the front end of the system is all differential.

After attenuation and scaling, the inputs are all reduced to a 10-mv level and are then amplified to a 10-volt level in the differential amplifier. At this point each input is compared to its respective high or low limit, which is set on a pin board. This feature will be used to transfer automatically the measuring cycle from 20 to 2 minutes during bad weather, occurring at any location along the 4.3-mile line. If a rain indicator, for example, should indicate rain for more than 3% of the time, then the comparator sends a signal to the programmer to put the system on a 2-minute scan rate.

The rules for the automatic operation of the 2- and 20-minute cycles are as follows (Fig. 14):

1. If the system is on a 2-minute cycle and sees an off limit on or before the ninth cycle, the out-of-limit detector will be set, the system will continue on 2-minute cycles, and the external indicator will stay on 2 minutes.

2. If the system is on 2-minute scans and does not see an off-limit signal by the ninth scan, it will reset the external indicator flip-flop at the time of change-over. The system will then go back to 20-minute cycles and no output from the indicator will be AND-ed with outputs coincident with the scans and cause the integrators to go to 20-minute time constants. The integrator time constants will remain correct during this change-over.

3. If the system is on a 20-minute scan and does not detect an off limit, the indicator flip-flop will not be set and the system will continue on 20-minute cycles.

4. If the system is on a 20-minute cycle and it detects an off-limit signal, it will set the indicator flip-flop after the last scan. The next nine scans will be 2 minutes apart. The output of the flip-flop AND-ed with outputs coincident with the scans will cause each integrator to go to the 2-minute constant during the first (02) scan. On this first 2-minute scan, the integrators will have the incorrect time constant. The output of the differential amplifier goes to an analog-to-digital converter whose output is in binary coded decimal form.

A digital clock which indicates minutes, hours, and days of the year is used to identify the time the data originated. Its output is scanned by the output control before the analog scanner goes into operation. During operation of the analog scanner the output control is connected to the analog-to-digital converter. The output control also inserts parity bits and flexowriter control signals. Its output goes to a remote paper-tape punch. The paper tape will be read one data period after its preparation. Buffering in the form of a paper-tape loop is provided between the punch and reader. After reading, the output is typed on a flexowriter. The flexowriter may be programmed to type all scans or to type only the integral 20-minute scans.

Fig. 15 shows the three systems tied together. It is necessary that an off limit at any station set the three stations on the same cycle or, manually setting any of the three systems to 2- or 20-minute cycles, set all the systems on the same 2- or 20-minute cycle. The outputs from

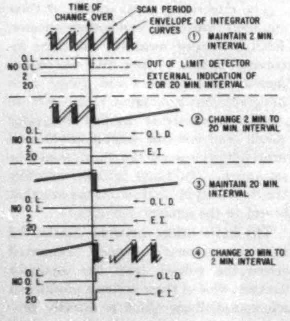

TIME OF CHANGE OVER
SCAN PERIOD
ENVELOPE OF INTEGRATOR CURVES
① MAINTAIN 2 MIN. INTERVAL
O.L.
NO O.L.
2
20
OUT OF LIMIT DETECTOR
EXTERNAL INDICATION OF 2 OR 20 MIN. INTERVAL
② CHANGE 2 MIN. TO 20 MIN. INTERVAL
O.L.D.
E.I.
③ MAINTAIN 20 MIN. INTERVAL
O.L.D.
E.I.
④ CHANGE 20 MIN. TO 2 MIN. INTERVAL
O.L.D.
E.I.

Fig. 14 (left). Analysis of use of 2- and 20-minute intervals

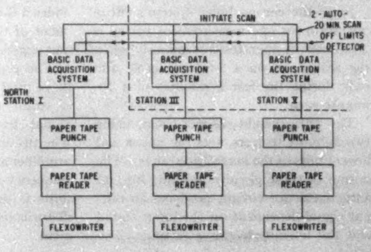

INITIATE SCAN
2-AUTO-20 MIN. SCAN OFF LIMITS DETECTOR
BASIC DATA ACQUISITION SYSTEM
BASIC DATA ACQUISITION SYSTEM
BASIC DATA ACQUISITION SYSTEM
NORTH STATION I
STATION III
STATION V
PAPER TAPE PUNCH
PAPER TAPE PUNCH
PAPER TAPE PUNCH
PAPER TAPE READER
PAPER TAPE READER
PAPER TAPE READER
FLEXOWRITER
FLEXOWRITER
FLEXOWRITER

Fig. 15 (right). Interconnection diagram of 310 data acquisition system

the three basic-data acquisition systems are all sent to the paper-tape punches at station I at the same time. Provision will be made for attenuation between stations III and V which are 2 and 4 miles respectively from station I.

At station I the output of the paper-tape punches will be read on three paper-tape readers. The flexowriters will type out the information from stations I, III, and V.

Summary of Data Processing

The processing and statistical analysis of the experimental data from Project EHV will be described in detail in a subsequent paper. Following is a brief summary of the planned operations:

1. The three paper tapes are converted every 2 weeks to magnetic tape, for processing through the IBM *705* computer in the Power Transformer Department of the General Electric Company in Pittsfield, Massachusetts.
2. All data are scaled, corrected for instrument systematic errors, and auxiliary quantities such as air density are computed. All resulting numbers are checked for validity (for instance, by noting if they fall within specified limits) and those not valid are printed out for visual inspection, to be reinserted later.
3. All "abnormal" conditions which had transferred the system from the 20-minute to the 2-minute cycle are printed out.
4. Averages, standard deviations, and frequency distributions are computed for all quantities and transferred to the master file.
5. The results of this run are combined with the corresponding results accumulated from the start of the test condition under study. Results (minimum and maximum values, averages, standard deviations, frequency distributions) are printed out in two columns under the headings "this run" and "start to date."
6. The effects of certain variables (such as voltage, humidity, aerosol density, etc.) on other variables (such as radio noise, corona loss, etc.) are determined by means of multiple-regression analysis.
7. An auxiliary program will allow the retrieval of all data for selected periods of time for detailed study of particularly interesting conditions; for instance, during severe rainstorms or sleet formation on the conductors.

Operating Experience to Date

As of July 1960 the equipment at the North Station had been installed and was under test. Initially, some of the transducers did not function properly so corrective action has been taken. Some of the analog equipment had changed characteristics at the higher temperatures of the final packaging. It dissipates 10 kw in the North Station building. This is currently being adjusted to correct for operation at these higher temperatures. Also, minor adjustments had to be made in the tape reader after a few hours of operation, and the pressure on the typewriter keys had to be reduced to increase the life of the ribbons. Spares are being purchased for the paper-tape-handling equipment and the typewriter.

The 0.6-mile short section of the EHV line from towers 1 to 3 was strung in September 1960 and energized in December 1960. The meteorological and mechanical data listed under station I in Table I were obtained beginning when the line was strung, and the electrical data were obtained as soon as the line was energized. The first conductor used was the Special (2.32-inch outside diameter). This has now been taken down and the whole 4.3-mile line is being restrung with two Plovers. The data from the line are being analyzed.

It is hoped that, by means of this unique data gathering and transmission system, the many and varied experimental data from Project EHV will be acquired and processed with maximum accuracy, speed, and economy, and transformed into meaningful conclusions which will help the electric utilities to meet the challenge of future EHV systems.

References

1. Project EHV, P. A. Abetti. *Electrical Engineering*, vol. 78, Aug. 1958, pp. 669–74.

2. Project EHV—Preliminary Corona Investigations, A. H. Foley, F. Olsen. *AIEE Transactions*, pt. III (*Power Apparatus and Systems*), vol. 79, June 1960, pp. 310–16.

3. Progress in Extra-High-Voltage Power Transmission, P. A. Abetti, S. B. Crary. *Ibid.*, vol. 78, June 1959, pp. 357–71.

4. Radio Noise Propagation and Attenuation Tests on Bonneville Power Administration McNary-Ross 345-Kv Line, G. E. Adams, T. W. Liao, M. G. Poland, F. J. Trebby. *Ibid.*, pp. 380–88.

5. The 400 Kv Research Installation Rheinau, Vol. I (book), G. Lesch, A. Rachel, editors. 400 Kv Research Society, Heidelberg, Germany, 1955.

Surface Electrical Failure in the Presence of Contaminants: The Inclined-Plane Liquid-Contaminant Test

K. N. MATHES
MEMBER AIEE

E. J. McGOWAN
NONMEMBER AIEE

ELECTRICAL FAILURE of the surface of insulating materials can occur in several ways. These include:

1. Air breakdown (flashover) which transfers to the adjacent insulation surface.
2. Thermal degradation of the surface (from arc interruption nearby, etc.) which temporarily or permanently damages the surface so that electrical failure occurs.
3. Partial volume breakdown, particularly of alminates, just beneath the surface which produces thermal or other degradation so that final failure ultimately occurs on the surface.
4. Surface degradation from electrical corona starting initially in limited surface areas exposed to voltage gradients high enough to produce localized air breakdown.
5. Progressive failure across the surface resulting from tiny arcs (scintillation) caused by conducting, or partially conducting, surface contamination.

Many tests have been devised to evaluate the resistance of insulation material to surface electrical failure.[1] However, relatively few such tests have become accepted as standard, and often even these are considered limited in scope or difficult to interpret in terms of service requirements. Much confusion has arisen because a test designed initially to evaluate one type of electrical surface failure has been indiscriminantly used to evaluate another, or even "all," types. In special cases, limited correlation between two tests has been shown,[2] but so many exceptions exist that correlation must be proved.

In recent years, the increased use of plastics and elastomers as insulation in

---

Paper 61-21, recommended by the AIEE Electrical Insulation Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 6, 1960; made available for printing November 15, 1960.

K. N. Mathes and E. J. McGowan are both with the General Electric Company, Schenectady, N. Y.

Table I. Comparison of Tracking Tests

Test	Nature	End-Point Criterion
Comparative tracking index of solid insulating materials under moist conditions[3]	Falling drops of 0.1% ammonium-chloride solution between brass electrodes held against surface	*Voltage* at which failure occurs after 50 20-millimeter drops have been allowed to fall
British "Admiralty," German *DIN 53480*, and other drop tracking tests	Falling drops of salt, ammonium-chloride, etc., solutions between electrodes on surface	*Time* (number of drops) to cause failure at specified voltage
Dust-fog tracking test (ASTM* suggested method)	Powdered, solid contaminant (salt, clay, fiber, etc.) dusted at intervals with continuous water mist	*Time* to electrical failure at constant voltage by tracking over the surface or breakdown through eroded volume

* American Society for Testing Materials.

areas where heavy surface contamination occurs, both indoors and out, has resulted in the development of tests to investigate the failure of materials under such conditions. Three such tests which have been intensively investigated are listed in Table I. Of these, the International Electrotechnical Commission (IEC) test for comparative tracking index has the longest history. The contaminant drop test seems to have appeared first in Switzerland (Ciba Company, Inc.?), about 1930 and in the United States about 1935 as the "Detroit Edison tracking" test. It was picked up in England by Metropolitan Vickers and further modified there as the "Admiralty" test. Various versions also developed in continental Europe. Finally, under the direction of P. D. Poppe of Oslo, Norway, valuable modifications were made, and the test procedure was published by the IEC.[3] The Norwegian modification introduced the important concept of voltage rather than just time as the principal end point of the test and results are obtained in a relatively short time. Subsequent investigation[4,5] has confirmed the critical importance of voltage in such tests. Unfortunately the IEC test is not sufficiently severe to evaluate the best materials available and the top test voltage has been extended from 600 to 1,000 volts, or even higher, in work undertaken in England, the Soviet Union, and the United States. Some of this work[5] has shown that above about 750 volts, instability, and sometimes electrode flashover, occur, which cause considerable difficulty. Consequently, the best materials cannot yet be determined by the IEC test.

A "dust-fog" tracking test, originally described by Starr and Albright,[6] evaluates the track resistance of very superior materials designed for severe outdoor service. In this dust-fog test, a constant voltage is applied across the surface of the sample which is periodically dusted with contaminant and kept wet with a water mist. Without question, this test has been of great value in classifying track-resistant insulation materials even though considerable variability has been encountered.[7] The long test time, sometimes 500 hours or more, is also a serious limitation, although recent work by Sommerman[2] indicates that the dust-fog test time can be reduced by about 10 to 1 with changes in the test parameters.

Recognizing the necessity of achieving a rapid, reproducible test, work was started late in 1958 to investigate the fundamental conditions and variables involved in surface tracking. The inclined-plane liquid-contaminant tracking test has been developed in this investigation.

Ambient Conditions for Surface Tracking

Field and laboratory investigations have shown that tracking occurs when contaminant deposited on the insulation surface becomes conducting in the presence of moisture from condensation or rainfall. Previous work[8] has shown that conductivity from surface contamination arises primarily from carbonaceous material (soot, etc.), fibers (lint, etc.), and ionizable contaminants such as inorganic salts. Other materials such as oxidized oil may tend to promote cohesion and adherence of the contaminant to the surface.

The key importance of moisture should be recognized. Even in the case of carbonaceous residues, some moisture, and perhaps ionizable contaminant, is often needed to provide electrical continuity between the individual carbon or graphite particles. Moisture is needed also to dissolve ionizable contaminants. Wet films proceeding over the insulation surface can spread conducting contaminant, which, in some cases, may be washed off structural surfaces and spread over the insulation surface. Examination of insulation surfaces tracked under natural conditions has indicated in many cases that moisture tends to collect and run down the *under* surface of the insulator to cause tracking, just below and, particularly, just above metal inserts or connections on the surface.

Many other factors are undoubtedly involved in surface tracking, even if indirectly. Outdoor exposure can produce surface degradation. The development of tiny surface fissures which promote track-

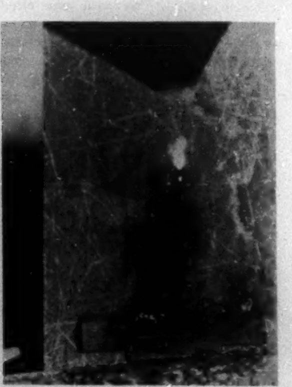


Fig. 1 (left). Test sample with electrodes attached

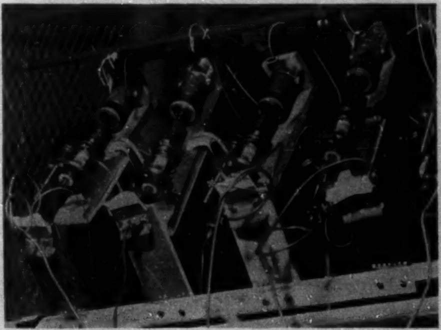


Fig. 2 (right). Test sample mounted in assembly at 45-degree angle

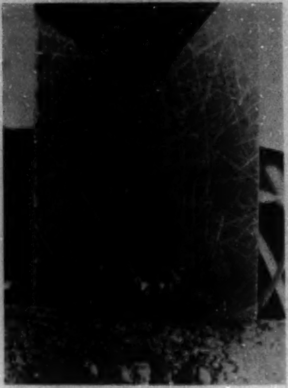


Fig. 3. Scintillation above bottom electrode (magnified)

ing, perhaps by trapping contaminant, have been described in a private communication from Dr. J. H. Mason, Electrical Research Association, Leatherhead, England. Chemical attack on the surface by deposited contaminant can occur and may be accelerated by the combination of moisture and sunlight. Oily residues are known to decompose in the presence of sunlight and air (oxygen) to give products like wetting agents which lower the contact angle of water on the insulation surface and increase the tendency of water films to spread. Chemical attack on the surface can also influence the spreading of moisture films containing ionizable contaminant. The geometry, position, and location of the insulator may greatly influence the nature and amount of contaminant deposited or collected on the surface.

The nature and severity of naturally occurring contamination is of great interest. Under the two major classifications, indoor and outdoor, many subclassifications are possible:

Outdoor Contamination

1. Sea coast: Severe salt encrustation from salt fogs, etc.
2. Industrial: Stack gas and other air-borne chemical contaminant.
3. Urban: Similar to, but less severe than, industrial-type contamination.
4. Rural: Rain, snow, and ice, but with limited dirt, etc.

Indoor Contamination

1. Heavy industrial: Heavy chemical contamination in cement mills, fertilizer plants, smelting and ore-reducing plants, chemical plants, dairies, etc.
2. Light industrial: Light chemical and lint contamination in power plants (fly ash), general industrial plants, etc.
3. Domestic: Dust, lint, etc., in "clean" industry, homes, etc.

Outdoor moisture exposure includes snow and ice in addition to rain and condensation. (In the United States, "outdoor" usually means fully exposed to the elements, but in Europe it may include enclosure in a roofed but unheated shed.) Melting snow may deposit trapped contaminants whereas severe rainstorms may wash away contaminants and clean the insulation surfaces. In outdoor applications, the concentration of soluble contaminant in the moisture may be quite low, while indoor contamination may build up slowly to considerable volume before condensation occurs to produce a quite concentrated solution. As will be shown later, electrostatic forces may blow contaminant solution off the insulation surface. Service conditions are highly variable, and it cannot be concluded that outdoor exposure is necessarily more severe than some indoor exposures. Field observation has indicated that the scintillation which causes tracking occurs for relatively short periods at rather widely spaced intervals of time, usually at the beginning, though sometimes at the end, of a "wetting" period.

Inclined-Plane Liquid-Contaminant Tracking Test

Recognizing that it is impossible in one laboratory test to produce all of the conditions in all the variation described in the previous section, it was decided to construct a test in which the principal variables involved in tracking could be introduced and each precisely controlled. Then, as different conditions needed to be investigated (i.e., different contaminants), they could be introduced one at a time. Every available source of background information was used and desirable features of previous tests were selected, combined, and used. The resulting procedure is called the "inclined-plane liquid-contaminant tracking test."

A sample with at least one flat face 2 by 5 inches [5 by 12.5 cm (centimeters)], preferably 1/4 inch thick, is mounted with the flat face on the under side at a 45-degree angle to the horizontal. Stainless-sheet-steel electrodes are fastened to the surface as shown in Fig. 1. Layers of filter paper are placed under the top electrode which has a hole in the bottom to permit exit flow of contaminant fed through a flexible plastic tube to the filter paper. The bottom electrode has a series of teeth bent to make spring contact with the sample surface with adjacent spaces to permit passage of contaminant without a damming effect. Contaminant fed to the top electrode flows down the under surface of the sample to and through the lower electrode. The arbitrary 45-degree mounting angle, shown in Fig. 2, can be varied but was selected here to give reasonably slow flow of contaminant without permitting it to drop off the under face.

Fig. 4. Typical tracking failure

When voltage is applied between the electrodes, electrostatic forces spread the contaminant stream into a delta shape—thick at the top and thin at the bottom. Within a critical range of voltage, the thin film of liquid contaminant boils and is otherwise disrupted so that discontinuities occur which produce scintillation primarily just above the bottom electrode, Fig. 3, where the liquid film is thinnest.

With proper test parameters, relatively steady and continuous scintillation can be produced. A series resistance improves the stability and also provides protection in case of sample failure. As voltage is increased, the scintillation current is increased which increases the severity of the test. Tracking failure of the insulation surface occurs (Fig. 4) when a critical voltage for each kind of material is reached. As in the IEC test, voltage is the critical variable most suitable as a comparison criterion.

Significantly, scintillation is maintained only when the voltage is varied through a

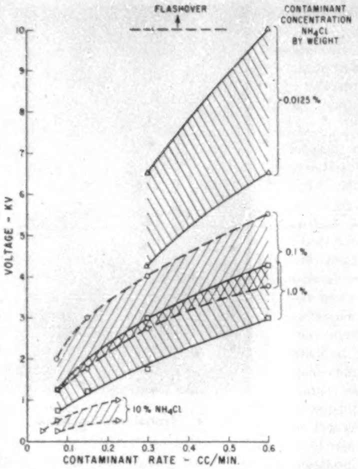


Fig. 5 (left). Range of voltage to produce scintillation for four different contaminant concentrations (electrode spacing: 2 inches, 5 cm)

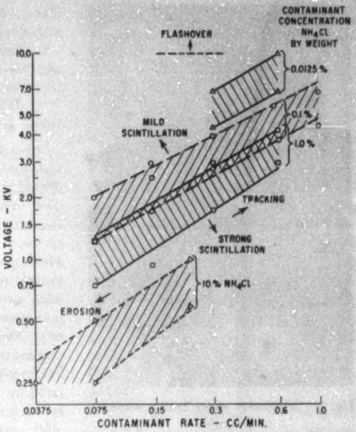


Fig. 6 (right). Data of Fig. 5 replotted on log-log scale

limited range. At voltages below the critical range, the current is conducted in the contaminant stream without scintillation. At voltages above the critical range, excessive boiling and electrostatic forces throw contaminant off the surface so that scintillation can occur only erratically and intermittently. However, by increasing the contaminant flow, the range of voltage at which scintillation occurs can also be increased, as shown in Fig. 5 for four concentrations of ammonium chloride. The very critical nature of contaminant concentration and the limited voltage range over which scintillation occurs demonstrate forcefully the importance of careful control of the test variables if a reproducible test is to be obtained. Moreover, it is apparent from these results that the right combination of conditions in service to produce damaging scintillation and resultant tracking can occur only by chance and very occasionally, which explains why tracking usually develops quite slowly in the field. Equally important, it can be seen that by producing steady scintillation under optimum conditions, tremendous acceleration can be realistically achieved in the laboratory.

By using only a liquid contaminant (electrolyte in solution), it is possible to obtain a far more precise control of contaminant application than is possible if a solid dust type of contaminant is also used. The function of the dust is primarily to provide a source of electrolyte and to aid in wetting the surface. The same effect is achieved with electrolyte in water solution applied in a controlled fashion, particularly if a wetting agent is added. However, it is difficult to decide what type and concentration of contaminant should be used. A wide variety of contaminants has been investigated. With some, like sulfuric acid, scintillation is relatively difficult to obtain. Others, such as sodium hydroxide, may attack certain insulation surfaces but can be used if this specific effect needs to be evaluated. Salts at equivalent solution conductivities in general produce tracking in most materials at nearly the same voltage levels, although erosion may differ as will be described later. Ammonium chloride has the advantage over sodium chloride that it does not tend to build a deposit on the sample during tests. Consequently, ammonium chloride has been used for much of the work to be described.

As shown in Fig. 5, at selected rates of contaminant application, a range of voltages for effective scintillation can be selected, depending upon the contaminant concentration. The data of Fig. 5 are replotted on a log-log scale in Fig. 6 to show that a power relationship is involved. It is perhaps even more important to determine how such changes in contaminant concentration will affect the tracking characteristics of materials and three typical materials were chosen for comparative test: *XXX* paper phenolic laminate (electrical grade), *GMG* Melamine glass-fabric laminate, and a track-resistant polyester glass-mat laminate.

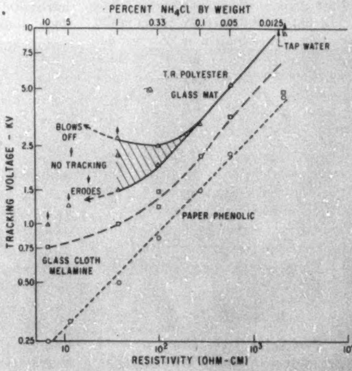


Fig. 7. Tracking voltage versus conductivity of contaminant for three different materials (contaminant rate adjusted for optimum scintillation)

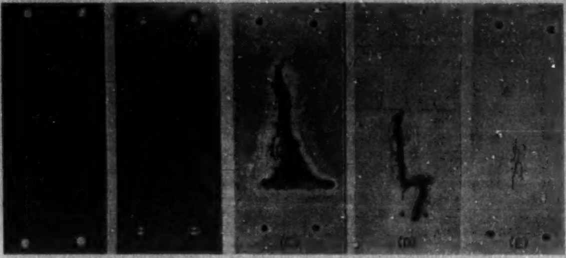


Fig. 8. Typical tracking failure at different concentrations of ammonium chloride

Sample	Concentration, %	Tracking Voltage
(A) XXXP	1.0	0.5
(B) XXXP	0.1	1.5
(C) GMG	1.0	1.0
(D) GMG	0.1	2.25
(E) GMG	0.0125	4.0

Table II. Comparison of Tracking Voltage, Inclined-Plane Liquid-Contaminant Tracking Test*

Material	Range† of Initial Tracking Voltage, Kv	Rate of Contaminant Application, Cm³/Minute
Molded wood-flour phenolic	1.0–1.5	0.075
Molded mineral-filled phenolic	1.25–1.75	0.075
Paper-phenolic laminate *XXX*	1.5	0.075
Silicone rubber	1.0–3.75	0.075–0.30
Glass-mat polyester laminate	2.0–3.25	0.15–0.30
Melamine-glass laminate *GMG*	2.25	0.15
Polystyrene	2.5	0.15
Polyethylene	2.75	0.15
Polyacetal (Delrin)‡	3.50	0.30
Polypropylene	3.75	0.30
Filled butyl rubber Hy Bute *60*)§	6.0¶ (flame failure)	0.90
Polymethyl methacrylate	6.0 (only rare flame failure)	0.90
Porcelain	6.0–8.0 (cracks or chips)	0.90–1.5
Polytetrafluoroethylene (Teflon)	7.0 (only occasional flame failure)	1.5

* 0.1% ammonium chloride.

† The values given cover the range of specific materials tested. Values for other compositions of the same basic materials may be outside the range given.

‡ Registered trade-mark of E. I. du Pont de Nemours & Company, Inc.

§ Registered trade-mark of the General Electric Company.

¶ At 6.0 kv and above, test conditions appear unrealistically severe and materials may be considered nontracking above this voltage.

The voltage at which tracking could be initiated in one hour is plotted as a function of contaminant resistivity in Fig. 7. (Contaminant concentration is plotted at the top.) Over very wide limits this log-log relationship for paper-phenolic laminate is a straight line with a slope of 1/2, indicating a power function which is probably directly related to energy input to the sample. The other two materials show a similar straight-line relationship only at lower contaminant concentrations. The tracking obtained with *XXX* and *GMG* laminates at lower voltages and higher contaminant concentrations are shown in Fig. 8. Under such conditions burning and erosion are emphasized. However, at a 0.1% concentration of ammonium chloride tracking appears to be the predominant factor, and this concentration with 0.02% nonionic wetting agent (Triton* *X-100*) added has been adopted for much of the work to date.

Table II shows a comparison of tracking voltages for various materials while Table III compares the dust-fog and inclined-plane tests.

TEST PROCEDURE

Since both voltage and rate of contaminant application need to be varied, it is desirable to develop a program which can be adopted as a standardized test procedure. (Unfortunately, standardization must necessarily decrease flexibility.) A method is being developed for presentation to the ASTM; the approach used is explained in the following. Steady 60-cycle voltage is applied between the top and bottom electrodes of the test sample after the contaminant flow has stabilized. The initial value of the voltage is selected by test and experience (described later) and then, after a selected period of time, is increased in 250-volt steps until progressive tracking starts, usually up from the bottom electrode. The test parameters are selected from Table IV.

The starting voltage of the test is selected by preliminary trial or experience so that tracking failure does not occur on the first voltage step (such results are discarded). It is usually necessary to use no more than four voltage steps to obtain tracking failure so that test time seldom exceeds 4 hours if 1-hour voltage steps are used. The test variability in itself appears to be limited to plus or minus one 250-volt step, but superimposed sample variability sometimes gives a wider range of test results for a specific material.

EROSION TESTS

The selection of time for test at each voltage is arbitrary. One hour at each voltage step has been selected for much of the work. At 500 volts below the "1-hour" tracking voltage, erosion may occur slowly above the bottom electrode, but tracking progresses very slowly, if at all. Voltage is a critical variable for tracking but time also must be considered in erosion studies.

A presentation of the considerable study of erosion would necessitate a complete paper in itself but typical results are given in Table V. The importance of the nature of the contaminant is very evident.

Both penetration and removed volume have been used to measure erosion. A dial micrometer depth gage with a 1/8-

Table III. Comparison of Test Results for ASTM Suggested Dust-Fog and Inclined-Plane Liquid-Contaminant Tracking Tests

Material	ASTM Dust-Fog Test* (Hours)		Inclined-Plane Test (Kv)	
	Average	Range of Values	Average	Range of Values
GSG (silicone-glass)	3.9	0.1–65	1.5†	1.25–1.5
PEG X1660‡	100	40–220	2.0	1.75–2.0
PEG X1494J‡	326+	95–410 (740E)§	3.0	2.75–3.25
Polymethylmethacrylate	167+ (E)	73(E)–350+(E)	Does not track up to 6.0 kv	

* Results reported by G. M. L. Sommerman.[7]

† Tracking occurred at values shown. Complete failure occurred at 3.0 kv.

‡ Material designation obtained from samples distributed for ASTM round robin test.

§ E = failed by erosion rather than tracking.

---

* Registered trade-mark, American Cyanamide Company.

Table IV. Test Parameters

Voltage Range, Kv	Rate of Contaminant* Application: Cubic Centimeters Per Minute	Rate of Contaminant* Application: Strokes Per Minute	Series Resistance, Ohms
1.00–2.00	0.675	1	1,000
2.25–2.75	0.15	2	10,000
3.00–3.75	0.30	4	50,000
4.00–5.00	0.60	8	50,000
5.25–6.00	0.90	12	50,000
6.25–8.00	1.50	12 (rate adjusted)	50,000

* A 0.1% solution of ammonium chloride with 0.02% Triton *X-100* in distilled water fed from an automatic repeating pipette delivering 0.075 cm^3 (cubic centimeter) per stroke. Samples which absorb moisture should be soaked in the contaminant before test.

At the lowest voltage range, 1.00–2.00 kv, scintillation is very critical and the rate of contaminant application may need to be varied to suit the characteristics of different materials. At the very lowest test voltages, the action of the pipette may be slowed to one stroke every 1.5, 2, or even 3 minutes.

Table V. Erosion Resulting From Surface Discharges, Inclined-Plane Liquid-Contaminant Tracking Test*

	Penetration, Inches		Eroded Volume, Cm^3	
Material	NH_4Cl	Na_2CO_3	NH_4Cl	Na_2CO_3
Filled butyl rubber	0.056 (0.036–0.084)		0.019 (0.012–0.028)	
Track-resistant polyester glass-mat laminate	0.032 (0.026–0.039)	0.103 (0.101–0.104)	0.017 (0.011–0.020)	0.339 (0.298–0.388)
Polymethylmethacrylate	0.029 (0.015–0.038)	0.080† (0.077–0.084)	0.012 (0.008–0.016)	0.020† (0.018–0.024)
Polytetrafluoroethylene (Teflon)	0.001‡ (0.001–0.002‡)	0.008 (0.006–0.009)	§	§

* In this test 2 kv was applied for 16 hours. The contaminant was applied at a rate of 0.15 cm^3 per minute. A 0.1% solution of ammonium chloride (NH_4Cl) in distilled water with 0.02% nonionic wetting agent (Triton *X-100*) was used. The concentration of sodium carbonate (Na_2CO_3) was adjusted to give a resistivity at 23 degrees centigrade equivalent to that for 0.1% NH_4Cl (282 ohm-cm).

† Tends to catch fire and burn (burned samples are discarded).

‡ Approximate.

§ Too small to measure.

inch-diameter long tapered probe was used to measure depth, while eroded volume was measured by pressing a putty-like compound (J-M Dux seal) into the cavity. The surface was scraped even with the original surface and the putty remaining in the eroded cavity was removed and weighed, and its density determined to permit calculation of the volume.

Conclusions

With the inclined-plane liquid-contaminant tracking test, the "wet" tracking resistance of a wide variety of materials can be evaluated in about 4 hours. The test results appear to be quite reproducible, apparently because the test parameters can be much more carefully controlled than in previous tests. The *voltage* end point of the test seems more capable of interpretation in terms of service requirements than *time* to failure; in service, the scintillation, which causes tracking failure, actually appears to occur only intermittently for relatively short and indeterminate lengths of time. The tracking voltage, as determined by the inclined-plane liquid-contaminant test, needs to be related to safe operating voltage by field test and such correlation seems theoretically possible if the field conditions can be defined. Without such correlation, the tracking voltage can be used as a comparative index (as with the IEC test[3]).

Moreover, it must be recognized that field conditions vary in many ways. Perhaps one of the greatest advantages of the inclined-plane test results from the ease with which special conditions can be represented by varying the nature, concentration, and amount of contaminant used. However, for a "standard" test, it is desirable to select one contaminant to provide an over-all basis for comparison. Both sodium chloride (common salt) and ammonium chloride have been considered and give comparable tracking voltage results for a number of materials. Ammonium chloride has been selected because it is easier to use, i.e., it does not tend to produce solid deposits.

Much can be done to improve the understanding of the tracking characteristics of insulating materials and their relationship to a variety of field conditions. For example, sunlight appears to have a considerable effect on various materials but has been evaluated so far in only a very limited way. Historically, surface tracking failure rather than erosion has been emphasized, although erosion resistance of materials varies greatly, and, particularly for those materials resistant to tracking failure, erosion may be more important. Erosion seems to depend much more on the nature of the contaminant than does tracking. The inclined-plane liquid-contaminant test can be adapted easily to the study of erosion under a wide variety of conditions.

Considerable progress has been made during the last few years in studying the effect of electrical stress on contaminated insulation surfaces. Such progress seems destined to continue at an accelerated rate to the benefit of the electrical industry. The wide interest in electrical tracking both in the United States and in many other countries provides important cross-stimulation which greatly aids in promoting these developments. A new proposal like the inclined-plane liquid-contaminant tracking test must be evaluated in different laboratories and undoubtedly will be further developed in such studies. The apparent reproducibilities, and rapidity with which the track resistance and erosion of a wide range of materials can be evaluated, should aid acceptance. The adaptability of the inclined-plane liquid-contaminant tracking test to a variety of conditions significant in service is perhaps one of the most important advantages which needs the confirmation of field experience.

References

1. TRACKING IN SOLID INSULATING MATERIALS: A CRITICAL RESUMÉ, **V. E. Yarsley, W. J. Grant.** *Report L/T 177*, British Electrical and Allied Industries Research Association, Leatherhead, England, 1948.

2. ELECTRICAL TRACKING RESISTANCE OF POLYMERS, **G. M. L. Sommerman.** *AIEE Transactions*, pt. III (*Power Apparatus and Systems*), vol. 79, Dec. 1960, pp. 969–75.

3. RECOMMENDED METHOD FOR DETERMINING THE COMPARATIVE TRACKING INDEX OF SOLID INSULATING MATERIALS UNDER MOIST CONDITIONS. *Publication no. 112*, International Electrotechnical Commission, Geneva, Switzerland, 1959.

4. DETERMINATION DE LA RESISTANCE AU CHEMINEMENT DE MATERIAUX ISOLANTS ORGANIQUES. *Bulletin no. 13*, Association Suisse des Electriciens, Zurich, Switzerland, 1959.

5. THE IEC COMPARATIVE INDEX TRACKING TEST, **K. N. Mathes.** Symposium on Tracking Resistance, American Society for Testing Materials, Cincinnati, Ohio, Mar. 1960.

6. TRACKING RESISTANCE TEST METHODS, **W. T. Starr, M. Albright.** *AIEE Transactions*, pt. III (*Power Apparatus and Systems*), vol. 75, June 1956, pp. 441–53.

7. THE ASTM ROUND ROBIN ON THE DUST FOG TRACKING TEST, **G. M. L. Sommerman.** Symposium on Tracking Resistance, American Society for Testing Materials, Mar. 1960.

8. METHODS FOR DETERMINING THE EFFECTS OF CONTAMINANTS ON ELECTRICAL INSULATION, **K. N. Mathes, L. E. Sieffert, H. P. Walker, R. H. Lindsay.** *AIEE Transactions*, vol. 68, pt. I, 1949, pp. 113–19.

◆

Discussion

Karl Lerstrup (Laur. Knudsen Electrical Manufacturing Company, Ltd., Copenhagen, Denmark): With the appearance of this paper a new light is thrown on an old problem, but rather than simplifying the issue it reveals a complexity that invalidates most of the criteria so far advocated for classification of insulating materials according to their tendency to develop tracking.

The importance of voltage was generally recognized with the acceptance in 1959 of IEC Publication *112* (reference 3 of the paper), but in the present paper new vistas are opened on the significance of the conductivity of the contaminant.

The laboratories of the company with which I am associated have used, since 1949, a method similar to the IEC test, based on voltage as the criterion, but with a standardized conductivity (360 ohm-cm) for the evaluation of phenolic moldings (only). With this method, it has been possible to achieve a consistent classification to ±10 volts for this type of insulating material, a consistency that was unattainable without a close control of conductivity.

The present paper takes an important step further and demonstrates (see Figs. 8(B) and (E)) that the same type of breakdown can be produced on melamine as on phenolic, but the conductivities of the contaminants must differ by at least a factor of ten. This is most significant for it points the way for further investigations in a direction that promises a better understanding of the factors significant for the proper evaluation of surface insulation quality. However, while the new approach may prove to be a valuable tool in the hands of the conscientious manufacturer, it is still a long way from anything that might justify any form of requirement in a national or international standard.

G. W. Heumann (General Electric Company, Schenectady, N. Y.): Failure of insulating materials along their surface between nonarcing current-carrying parts of electric devices is considered by the design engineer in choosing proper creepage distances. For some electrical products, U.S. standards specify creepage distances. For other products, only a dielectric test is prescribed to insure adequate spacing of live parts. In either case the choice of proper creepage distances and dielectric test voltages is a factor in, based on experience, the effect of contaminants as usually found in industrial atmospheres. No attempt has been made to establish standards on ambient conditions, and extreme cases of contamination are taken care of largely by special design based on operating experience in particular locations.

Creepage distances and dielectric test voltages in U.S. standards are graded by rated voltage of the device without regard to properties of the insulating material along the surface of which a breakdown would occur. There has been a feeling among designers that the "tracking" properties of various materials differ from each other. However, we felt that we do not know enough about the physics of surface failure to classify insulating materials for the purpose of establishing different classes of creepage distances.

In Europe it was felt that, for economy in the utilization of insulating materials, they should be classified, and reduced creepage distances should be permitted for materials with favorable tracking properties. As a result IEC Publication *112* (see reference 3 of the paper) establishes a "tracking index" for grading materials. Several countries have accepted the IEC index as a design guide. However, it was never accepted in the United States because we felt that the IEC test is oversimplified and does not yield reliable and consistent classification data. We do not feel that we know enough about the mechanism of surface breakdown to make any tracking index a guide for design.

It is very interesting to note that Dr. Lerstrup, who is active in IEC work and who strongly supported the IEC tracking index as a design guide, now acknowledges that, as demonstrated in the paper, surface tracking is a more complex phenomenon than the IEC test implies. The authors have made a valuable contribution to our understanding of tracking phenomena. It is hoped that further work will be done, and that some day we can take classification of materials according to tracking properties into consideration when establishing creepage distances. However, with our present knowledge of materials, I believe that we have been correct in our contention that U.S. standards should not yet establish creepage distance classes for different kinds of materials.

W. Kaufmann (Siemens-Schuckert Company, Amberg, Germany): The new testing method resembles in many respects the "drop test" described in the German VDE Standard[1] and in IEC Publication *112*.[2] In the inclined-plane test a liquid contaminant is used which is, in fact, the same type as specified for the drop test. Also, scintillations are created by adjustment of voltage, concentration, and liquid flow; however, this can be done in a much simpler manner with the drop-test method. Differences in the phenomena accompanying tracking can be explained by the much greater electrode spacing (5 cm as compared with 4 millimeters for the drop test) and the correspondingly high voltage (up to 8 kv as compared with 600 volts). This testing method described is, like several other American tests,[3] a specific high-voltage test, whereas the drop test is specifically restricted to low-voltage applications.

In order to establish creepage distances for low-voltage devices, the authors propose to divide insulating materials into classes with which the various creepage distances are to be co-ordinated. For the establishment of tracking properties of insulating materials, a test utilizing a low voltage is preferable to one based on the use of a much higher voltage. It is better practice to base a test of insulating materials on a voltage of the same order of magnitude as the voltage rating of low-voltage switching devices. If this is not done, and if significantly greater distances and significantly higher voltages are employed for the testing, phenomena will occur which no longer correspond to the stresses occurring in actual practice.

For these reasons I do not believe that the Mathes-McGowan test is an acceptable replacement for the drop test. Furthermore, the paper cites only a limited number of tests on different insulating materials, and these are all laminated or molded compounds. If a new method is to be prescribed for testing materials, it is first necessary to make extensive tests in a number of laboratories. Drop tests were conducted for over 10 years in various testing laboratories, and today it can be stated that such tests make it possible to establish distinct classes of insulating materials for low-voltage apparatus.

If a tracking index is to be used for establishing creepage distances for low-voltage devices, this index should be tested by a method which had been proved by many years of use and which is not at all inferior to the proposed method. However, it is not intended to deprecate the importance of the method described in the paper for further studies of tracking phenomena, particularly in relation to high-voltage apparatus.

REFERENCES

1. *VDE Standard 0303*, Verband Deutscher Elektrotechniker, Wuppertal-Elberfelt, Germany.

2. See reference 3 of the paper.

3. DRY LINEAR TRACK TEST AND CIRCULAR TRACK TEST. *ASTM Standard D495*, American Society for Testing Materials, Philadelphia, Pa.

G. M. L. Sommerman (Westinghouse Electric Corporation, Pittsburgh, Pa.): The authors have presented an interesting new inclined-plane test procedure for evaluating the resistance of materials to electrical surface tracking under moist conditions. The idea of applying the contaminant to the under side of the test specimen has considerable merit; a similar scheme was employed in European tracking tests in which steam was allowed to condense on the under side of the specimen resting on knife edges. Principal advantages of the Mathes-McGowan procedure are: (1) test results are obtained in less than 5 hours on even the best materials, thus allowing rather quick screening of materials over a wide quality range; and (2) as the authors state, the flexibility of the test permits studies of erosion and other surface-degradation phenomena under a wide variety of conditions. Most other screening tests have been found to be limited to relatively weak materials.[1]

Many of the materials reported in this paper have been evaluated by the writer with a linearized accelerated version of the dust-fog test.[1,2] These include: polymethylmethacrylate, 33 hours; polyacetal, 35 hours; polypropylene, 40 hours; and polytetrafluoroethylene, 330 hours (the times indicated are those for the occurrence of complete tracking failures). It is interesting to compare these results with those reported in Tables II and III of the paper. The comparison indicates excellent agreement in general, with the following principal exceptions. Our data indicate that polystyrene is weaker than melamine glass cloth, 2.5 kv, rather than stronger, as reported in Table II. Also, our results indicate that methyl methacrylate is weaker than polypropylene and filled butyl rubber, rather than stronger as reported in Table II, and that the methyl methacrylate is weaker than polyester glass *X 1494DJ*, rather than stronger as reported in Table III.

One possible explanation is that the results in Tables II and III are in terms of the initial voltage to cause tracking rather than the voltage to cause the tracks to propagate a large distance, say 2 to 5 cm. I originally proposed that initial tracking voltages had merit in the dust-fog test[3] as a quick indication of relative quality of the weaker materials. However, as better materials are considered, it is no longer possible to correlate the final tracking life with the initial tracking voltage.

Another explanation of the above may lie in the difficulty encountered generally in trying to use the results of increasing-stress tests to predict life at a relatively low constant-stress condition. While the value of the voltage exponent n in the insulation-life equation may be similar for a number of materials, there are always a few materials for which the value of n is different. For materials of the latter type, the long-time behavior cannot be predicted accurately from short-time tests.

In order to compare the relative severities of these inclined-plane tests with those of the procedures, it would be helpful to know what values of scintillation currents and wattages were measured for the various contaminant concentrations and flow rates indicated in Figs. 6 and 7.

Further comment is requested on the authors' statement that, "The *voltage* end point seems more capable of interpretation in terms of service requirements than *time* to failure." I question this conclusion; also, I do not follow how it is supported by the next statement, "in service the scintillations . . . occur only intermittently for relatively short and indeterminate lengths of time." One reason for questioning the first statement is that the numerical value of the exponent in the voltage-versus-life characteristic may depend on the composition of the specimen, as noted above. Aside from this, would we feel safe in using a phenolic insulator in severe outdoor conditions at a voltage equal to one third of the relatively high voltage at which we could safely operate a large insulator made of polyacetal or polypropylene? The test results (voltages) in Table II are in this ratio. With regard to the second part of the statement, our observations on insulations tested outdoors under voltage indicate that scintillations occur a large number of times in any one year.[4] Thus in each successive year tracks will grow a certain amount on fairly weak materials and a certain amount of degradation will occur on even good materials. It would appear, therefore, that the expected outdoor service life of materials would be more directly related to the length of time the material can withstand a constant-stress type of wet tracking test such as our linearized accelerated dust-fog test. Eventually the latter test might be reserved for only those materials that the inclined-plane test indicate to be suitable for a given application after a general correlation of results from the two types of test is established.

REFERENCES

1. ELECTRICAL TRACKING RESISTANCE OF POLYMERS. II. RESULTS WITH LINEARIZED AND ACCELERATED TEST PROCEDURES, J. Chottiner, G. M. L. Sommerman. *Annual Report*, Conference on Electrical Insulation, National Research Council, Washington, D.C., 1960.

2. See reference 2 of the paper.

3. Discussion appended to SUGGESTED METHOD OF TESTING TRACKING RESISTANCE OF MATERIALS. "ASTM Standards on Electrical Insulating Materials," American Society for Testing Materials, Philadelphia, Pa., Sept. 1957, p. 641.

4. OUTDOOR INSULATION TESTING, R. T. Innis, R. F. Gieselman. *Report*, Westinghouse Electric Corporation, Bloomington, Ind., May 19, 1960.

R. J. Bronikowski (Line Material Industries, South Milwaukee, Wis.): This paper is both interesting and timely. A tracking test which can minimize the lengthy evaluation time of present tests is of potential value to all investigators. Would the authors describe the test setup in greater detail? Of particular interest are details of the electrode configurations and the test circuit parameters.

The performance of an insulating material after a period of natural or artificial weathering is frequently of greater significance than its performance when "new." The authors point out: "The effect of sunlight on various materials appears to have a considerable effect (on tracking characteristics) and has been evaluated so far in only a very limited way." In this limited evaluation, do test data from the inclined-plane tracking test show the expected reduction in "tracking voltage" after exposure of samples to artificial or natural weathering conditions?

K. N. Mathes and E. J. McGowan: In general reply to several of the discussions, it should be emphasized again that our paper has described a study of several parameters involved in the surface electrical failure of plastics and insulating materials using a new experimental approach called the inclined-plane liquid-contaminant tracking test. To date, no *specific* tracking test has been formally proposed by us although it appears possible that useful versions of the method can be developed ultimately for such purposes as materials comparison, specification, acceptance, and screening test. However, a number of variables must be fixed before the method can be used easily as a specific test. Unless a completely arbitrary selection is to be made, the choice of variables for a specific test should be based upon conditions the insulating materials will meet in service. As Dr. Lerstrup and Mr. Heumann have pointed out, such conditions need to be investigated to determine so far as possible the best way of setting limits on the test variables.

Special thanks are given to Mr. Heumann for his helpful suggestions and particularly for his interest in obtaining comments from a number of European engineers.

In specific reply to Mr. Kaufmann, it is agreed that test voltages and spacings should be representative of those in specific apparatus when the test is used to evaluate materials for such apparatus. The inclined-plane test has been used at voltages as low as 250 volts and with electrode spacings as small as 1/4 inch (0.65 cm). In a number of cases tracking voltage has been shown to be a linear function of electrode spacing, from 1/4 to 10 inches (see Fig. 9). This information, combined with oscillographic studies of the current waveform, indicates that the phenomena causing scintillation and tracking are probably much less dependent on total voltage than Mr. Kaufman believes. If information to the contrary is available, we would be very interested.

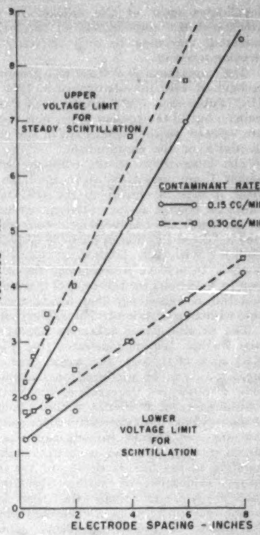


Fig. 9. Voltage ranges for scintillation CNH$_4$Cl

In contradiction to Mr. Kaufmann's comment, the inclined-plane test differs significantly from the VDE and IEC drop tests in that contaminant solution is continuously applied to the test surface at a *controlled* rate which is selected so as to be suitable for the voltage and solution conductivity used. It is not agreed that the test parameters can be more *easily* varied with the "drop" tests. However, it is recognized that useful information can be obtained by varying the parameters of the drop test as has been shown by a number of investigators including V. Siegel at the Physicalisch Technische Bundesanstalt in Brunswick, Germany, and N. Parkman at the British Electrical and Allied Industries Research Association in Leatherhead, England.

Unfortunately, even with such variations, the drop test apparently cannot be used to evaluate the tracking resistance of the better materials which are easily evaluated with the inclined-plane test. Moreover, it is difficult to evaluate erosion quantitatively with the drop tests or to separate erosion from tracking. Nevertheless, the drop tests may be useful in qualitatively comparing tracking characteristics of the poorer materials.

We consider the drop tests to be of limited usefulness, and perhaps even misleading, if used for establishing differentiation in design creepage distances of different materials in specific types of apparatus as proposed by Mr. Kaufmann. As pointed out by Dr. Lerstrup, in such specific applications many complex factors are involved which are not yet sufficiently correlated with the parameters of *any* of the tests now in use. It is hoped and expected that the flexibility of the inclined-plane test combined with much additional experience and the study of service conditions for specific apparatus may ultimately make possible the separation of materials into classes for which different electrode spacings for specific types of apparatus may be specified.

Dr. Sommerman's comments are particularly appreciated because of his considerable experience with tracking phenomena. He points out that the order of merit of different materials sometimes depends upon the type of test used. It is precisely this point which makes us reluctant to fix any one set of variables for the inclined-plane test since we have found also that differences between materials depend considerably upon the test parameters selected. Unlike many other tests, the inclined-plane test permits investigation of the effect of a wide range of variables. As Mr. Bronikowski suggests, outdoor weathering can adversely affect the tracking voltage. However, careful preparation of the sample surface, the removal of oil or wax, and soaking of the sample prior to test, reduces such differences. Experience is limited, but to date only a drop of 0.25 to 0.50 kv in tracking voltage seems to be brought about by ultraviolet weathering.

The selection of tracking voltage as a test end point and basis of comparison has been questioned by Dr. Sommerman. As pointed out in the paper, scintillation in service occurs at widely spaced time intervals for relatively short periods. The severity and frequency of scintillation in service also varies widely from time to time and place to place. Yet failures may occur under many such conditions. It seems more significant to determine the voltage stress below which tracking will never occur rather than the length of time-to-failure conditions. The ratio of failure voltage to which Dr. Sommerman refers depends upon the test parameters. For a track-resistant material as compared with a phenolic, it may be 2 to 1 or 5 to 1 or infinity. (At high contaminant concentrations the track-resistant material erodes rather than tracks; see Fig. 7.) Again the choice of parameters is important. Fig. 7 shows that phenolics cannot be used outdoors. In other tests phenolics have failed at voltage stresses as low as 100 volts per inch (750 volts over $7^1/_2$ inches).

With the dust-fog tracking test mentioned by Dr. Sommerman, the amount and concentration of contaminant is difficult to control and varies during test in a largely unpredictable fashion. The continuity, severity, and character of the scintillation consequently varies. Probably for part of the time tracking predominates and for part of the time primarily erosion takes place. When test parameters vary widely in this way, it may be necessary to use time to failure as the test end point hoping that a "smoothing out" of the test conditions will occur over long time intervals. Erosion, however, in contrast to tracking, is time-dependent as has been shown with the inclined-plane test in which erosion can be evaluated separately from tracking.

As mentioned in the paper, resistance to erosion may be even more important in the application of good track-resistant materials under severe exposure conditions. With the inclined-plane test, the erosion of different materials can be compared after about 16 hours of test at a single voltage by measuring the depth of penetration and the eroded volume. It is possible also to arrange electrodes so that time to failure by erosion may be measured.

Analog and Threshold Building Blocks for Variable-Radix Adders and Other Logic

J. SKLANSKY
MEMBER AIEE

Synopsis: A system of "TKM" logic, based on 3-state threshold devices, Kirchoff adders (i.e., summing amplifiers), and linear amplifiers is proposed. It is shown that TKM logic can realize any multiple-valued truth table, and that the length of the alphabet of truth values may be arbitrary. This logic is shown to be particularly well suited to the synthesis of variable-radix adders, in which alterations of the radix are obtained by adjustments of the gain or attenuation of one or more amplifiers. When TKM is used in conjunction with a proposed "symmetric analog coding" for the representation of numbers, the generation of $r-1$ complements and, as a consequence, the execution of r-ary subtraction and division, are facilitated. It is conjectured that the feasibility of variable-radix adders is a step toward the realization of adaptive computation at the hardware level.

THERE IS a tendency to assume that the realization of an m-valued system of logic requires m-state digital devices, just as the realization of conventional 2-valued logic is invariably obtained with 2-state digital devices. A system of r-valued logic (r arbitrary) is proposed herein in which the building blocks consist of analog devices and 3-state threshold devices. Specifically, the following idealized building blocks are proposed: (1) 3-state threshold devices, (2) Kirchoff adders (i.e., summing amplifiers), and (3) linear amplifiers. [Sometimes threshold devices (or "limiters") are classed as analog components.[1] However, in this paper they are regarded as extended forms of OR gates and AND gates, which fall in the category of digital components.]

The Kirchoff adders and linear amplifiers, of course, must be able to discriminate among r signal levels, r being the multiplicity of the truth values. In this sense the Kirchoff adders and linear amplifiers are r-state devices. However, these components have the distinction that their effective number of states (i.e., their allowable number of signal levels) is determined by their environment of circuitry and signal coding, rather than by their internal physical characteristics. Thus the same components may be used for a variety of radices or truth-value multiplicities.

The proposed logic system shows particular utility for the realization of variable-radix adders. In these adders the radix may be altered merely by adjusting the gain on certain amplifiers (or, equivalently, adjusting certain attenuation resistors). Other multiple-valued logic circuits are also realizable; in fact it is shown that the three building blocks can realize any r-valued truth table. (The value of r is usually limited by physical and engineering considerations, such as the components' electrical tolerances, the signal-to-noise ratio, and the desirability of matching the internal coding to the input or output equipment.)

The existence of variable-radix adders may lead to the operation of entire computers with variable-radix arithmetic. One advantage of such machines will be the possibility of adaptive radix selection, which is adaptive in the sense that the machine will choose the radix which minimizes a cost function of computation time and accuracy.

If "significant-digit arithmetic"[2] is used, the size of the radix can control the accuracy because the "significant" portion of a numerical result can be chosen with small round-off error when the radix is small, while a relatively large round-off error must be accepted when the radix is large. On the other hand, as will be demonstrated later, the computation time will be short when the radix is large,

---

Paper 60-805, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Atlantic City, N. J., June 19–24, 1960. Manuscript submitted March 22, 1960; made available for printing November 9, 1960.

J. SKLANSKY is with RCA Laboratories, Princeton, N. J.

The author wishes to acknowledge the helpful comments of many RCA staff members; in particular, S. Amarel, A. Lo, J. T. Wallmark, and R. O. Winder of RCA Laboratories, and S. Marcus of RCA Defense Electronic Products.

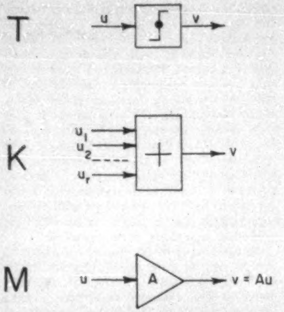
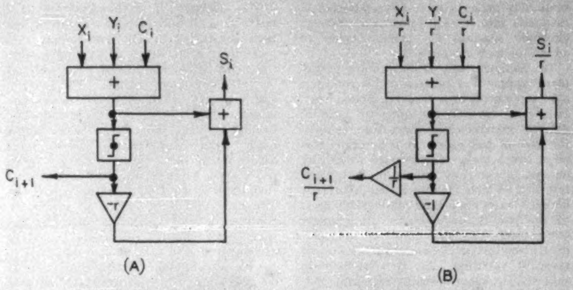


Fig. 1. Graphical symbols for T, K, and M. The letter A inside the triangular block represents the multiplying constant of the multiplier



(A) (B)

Fig. 2. TKM-realizations of an r-ary full adder. Standard SAC signals are utilized in the circuit of (A); normalized SAC signals in (B)

and long when the radix is small. Hence time and accuracy can be bartered by means of a radix adjustment.

The Logic System: TKM

The proposed logic system consists of a coding scheme and a set of basic building blocks.

The Coding Scheme

The coding scheme is directed toward the facilitation of addition and subtraction, and, as a result, arithmetic processes in general.

Suppose the number system has the radix r. Then, by this coding scheme a digit x has the following signal representation:

$$X = 2x - r + 1 \tag{1}$$

Thus if $r=5$, x will ordinarily range over 0, 1, 2, 3, 4. As a result, X will have the values -4, -2, 0, 2, 4, respectively.

Lower-case letters will be used to represent actual truth variables, while capital letters will represent the corresponding signals. For example, y will have the signal representation

$$Y = 2y - r + 1 \tag{2}$$

An important advantage of this coding scheme is that replacing X by $-X$ is equivalent to replacing x by $(r-1)-x$, thereby obtaining the $r-1$ complement merely through an inversion of polarity. (With the $r-1$ complement of a number, subtraction can be implemented simply by entering the complement into the adder, and adding 1 to the resulting sum.) Another advantage lies in the detection of "independent" and "dependent" carries for the addition process. (An "independent" carry is a carry whose value is independent of the carry brought forward from the preceding column. A "dependent" carry is dependent on the preceding carry, and is in fact equal to it.[2]) Consider, for example, the problem of adding 3 and 2 in the quinary system. Let $x=3$, $y=2$; then $X=2$, $Y=0$.

Now suppose X (and Y) are written as a summation in the form of a series of $+1$'s and -1's. Suppose the series is so chosen that the total number of terms is $r-1$, i.e., 4 in the present example. Under the proposed coding scheme this will always be possible. In fact, the series for X (or Y) can be obtained by noting that the number of $+1$'s is x (or y), and the number of -1's is $r-1-x$ (or $r-1-y$). Furthermore, for reasons which will become evident later, all the -1's are written to the left of the $+1$'s in the series expression for X, and the -1's to the right of the $+1$'s in the expression for Y; and each -1 or $+1$ is aligned in the expression for X over a corresponding -1 or $+1$ in the expression for Y. Thus, in the present example ($X=2$, $Y=0$),

$$X = -1+1+1+1 \tag{3A}$$

$$Y = +1+1-1-1 \tag{3B}$$

Note that the line of $+1$'s in equation 3(A) overlaps the line of $+1$'s in equation 3(B). The reader is asked to verify that under this condition of overlap the sum of x and y is necessarily greater than or equal to r, and consequently the carry resulting from the quinary column sum, $x+y$, is an independent carry of value 1.

If the line of -1's in equation 3(A) had overlapped the line of -1's in equation 3(B), the carry would again have been independent, but this time would have had the value 0. If there had been no overlap of -1's or $+1$'s, then the sum of x and y would have been 4 (or in general, $r-1$), so that the carry would then be dependent and equal to the carry from the preceding quinary column. Thus it is seen that

$$\begin{aligned} &X+Y>0 \text{ implies: carry}=1\\ &X+Y=0 \text{ implies: carry}=\text{preceding carry}\\ &X+Y<0 \text{ implies: carry}=0 \end{aligned} \tag{4}$$

This is an indication that the coding scheme will permit implementation of addition with a simple form of threshold device. Later this will be demonstrated with a specific adder network. (Relations 4 could also have been obtained through equations 1 and 2; however, equations 3(A) and 3(B) were used because they constitute a useful mental tool for visualizing the independent–dependent carry conditions.)

In the sequel, square brackets will be used to indicate the signal representation of a truth variable. Thus

$$[x] = X$$

This will be convenient for representing such quantities as $[x \text{ OR } y]$ or $[3x]$ without new capital-letter symbols being assigned.

The above coding scheme will be referred to as "SAC," for "symmetric analog coding."

The Building Blocks

The following three building blocks are postulated: 3-state *T*hreshold devices (T), *K*irchoff adders (K), and constant *M*ultipliers or linear amplifiers (M). Memory cells or timing devices of any sort are not included, thereby restricting the method to nonsequential logic. Precise definitions of T, K, and M are given in the following, and corresponding graphical symbols are shown in Fig. 1.

Definition of T: One input, u; one output, v.

$$v = \begin{cases} 1 \text{ if } u>0 \\ 0 \text{ if } u=0 \\ -1 \text{ if } u<0 \end{cases} \tag{5}$$

In an actual T, of course, the range of u over which $v=0$ will be a neighborhood around 0 rather than just the point 0. For the purposes of this discussion, however, it is convenient and no less general to replace the neighborhood by the point.

A convenient way of realizing a practical T is by means of two 2-state devices such as diodes. Often, however, only a single 2-state device will be necessary, because in many cases the input to T is never zero.

Definition of K: Any number of inputs, u_i; one output, v.

$$v=\sum_i u_i \tag{6}$$

where the summation is by arithmetic addition.

Definition of M: One input, u; one output, v.

$$v=Au \tag{7}$$

where A is some predetermined real constant: positive, negative, or zero.

The problem of realizing M with sufficient linearity and speed for modern very-high-speed computers is at present more difficult than the realization of T and K. New devices, however, are always appearing on the scene, and this situation may be changed in the future.

A circuit that has been realized solely by T, K, and M is referred to as a "TKM circuit." Although TKM logic does not necessarily involve the SAC coding scheme, the use of such a coding scheme will be assumed in this paper whenever a reference is made to TKM logic.

It should be realized that the functions T, K, and M do not necessarily correspond to distinguishable physical components. It is expected, actually, that many physical components for TKM circuits will incorporate two or more of these functions in various combinations. Examples are resistor-diode circuits combining T and K, and resistor-amplifier circuits combining K and M.

Applications

This section will describe adders and other computer organs realized by TKM logic. The adder networks seem particularly well mated with TKM.

Full Adder

A full adder is a 3-input 2-output adder. The inputs consist of two r-valued signals X_i and Y_i, and a 2-valued carry signal C_i; the outputs are an r-valued sum signal S_i and a 2-valued carry signal C_{i+1}. The relationship between these signals is determined by the following truth-variable inequalities:

$$s_i=\begin{cases} w_i \text{ if } w_i<r \\ w_i-r \text{ if } w_i\geq r \end{cases} \tag{8A}$$

where

$$w_i=x_i+y_i+c_i$$

$$c_{i+1}=\begin{cases} 0 \text{ if } w_i<r \\ 1 \text{ if } w_i\geq r \end{cases} \tag{8B}$$

The full adder is a basic circuit that can be used in a variety of complete adder circuits, such as the simple series adder, the simple iterated adder, and the independent–dependent carry adder.[3] Fig. 2(A) shows a suggested TKM realization of such an adder. In this circuit X_i, Y_i, and S_i are r-ary SAC representations of x_i, y_i, and s_i; while C_i and C_{i+1} are binary SAC representations of c_i and c_{i+1}. Thus,

$$X_i=2x_i-r+1 \tag{9A}$$

$$Y_i=2y_i-r+1 \tag{9B}$$

$$C_i=2c_i-1 \tag{9C}$$

This circuit works as follows: The 3-input Kirchoff adder produces the signal

$$2[(x_i+y_i+c_i)-r]+1 \tag{10}$$

Consider the case where the output carry, c_{i+1}, is 0. Under this condition, the bracketed quantity in equation 10 is ≤ -1. The entire signal, equation 10, is then ≤ -1, thereby producing -1 at the output of the threshold device. (This is the correct value of C_{i+1} corresponding to an output carry of 0.) The signal -1 is multiplied by $-r$ in the multiplier, thus feeding $+r$ to the lower input terminal of the 2-input Kirchoff adder. The upper input receives the signal, equation 10, thereby producing

$$2[x_i+y_i+c_i]-r+1$$

which is the corect SAC-representation of s_i when c_{i+1} is 0. The reader is asked to verify S_i and C_{i+1} for the case $c_{i+1}=1$.

Note that the radix, r, can be changed by adjusting the gain, $-r$, of the M. At the same time, of course, the coding of the input summands must be changed appropriately.

An advantage of this circuit is that the threshold device is called upon to produce only two values, rather than three. Hence T can be realized with only a single diode or other 2-state device. For such situations, however, the dot in the center of the zigzag symbol is still used so that the variety of symbols may be at a minimum.

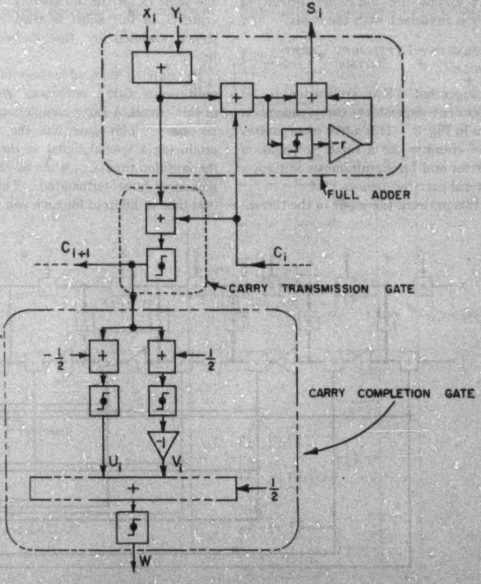


Fig. 3. A TKM-realization of an r-ary independent–dependent carry adder

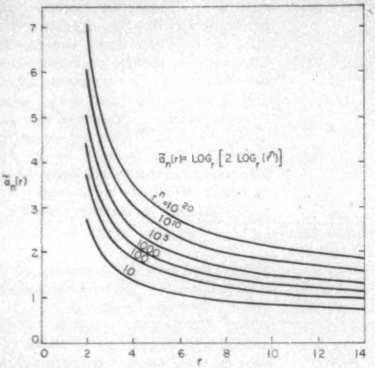


Fig. 4. The upper bound, $a_n(r)$, of the average longest sequence of carries of 0's or 1's as a function of the radix r, with the adder capacity, r^n, as a parameter

In order to implement a variable-radix adder it seems important to maintain an invariant dynamic range (or swing) at all points in the circuit as the radix changes. In this way the signal–noise discrimination will improve as the radix decreases. A full adder realizing a constant dynamic range is shown in Fig. 2(B). Here the signals and the multiplier gain of Fig. 2(A) have been divided by r. As a result the signal swing at the output of every T, K, or M except the multiplier producing C_{i+1}/r is invariant with the radix.

INDEPENDENT–DEPENDENT CARRY ADDER

A suggested TKM realization of an independent–dependent carry adder is shown in Fig. 3. This adder incorporates special circuitry for a rapid propagation of carries and for simultaneous initiation of several carry sequences.[3,4]

In this network the input to the threshold device inside the carry transmission gate (enclosed by dashed lines) is equal to C_i whenever $x_i+y_i=r-1$, since X_i+Y_i is then equal to zero. Under this condition the output of the threshold device, C_{i+1}, is equal to C_i. On the other hand, whenever x_i+y_i is less than or greater than $r-1$, the output of the threshold device is clearly -1 or $+1$, respectively, thereby producing the correct independent carry. (The coding for C_i, X_i, and Y_i is the same as for the full adder.) To generate S_i, a full adder is used. This is shown enclosed by the dashed-dotted line.

In order to take advantage of the simultaneous carry sequences generated in this circuit, a carry completion gate is necessary. This gate has the task of producing a special signal at the end of the addition time, i.e., when all the carry sequences have terminated. This addition time is different for each pair of summands, and is proportional to the length of the longest sequence of 0-carries or 1-carries. A gate fulfilling this function is shown enclosed by the dashed-dotted-dotted line in Fig. 3. This gate consists of two branches emanating from each carry point, all leading into a multi-input Kirchoff adder. Each pair of branches is in essence a biased-output full-wave rectifier producing the following signal:

$$U_i+V_i=\begin{cases}-2 \text{ if } C_{i+1}=0\\ 0 \text{ if } C_{i+1}=\pm 1\end{cases} \tag{11}$$

As a result the carry completion signal, W, is $+1$ if all the C_i's are ± 1 and -1 if any C_i is 0.

Hopefully, an adder of this kind can take advantage of high signal-to-noise ratios in present digital devices, and adapt them to analog use, thereby packing a greater density of information-processing into the same equipment. It is noted also that the number of inputs leading into any component as well as the number of output leads emanating from any device is quite low throughout the adder circuit; hence at least a partial implementation of this adder by the new threshold devices requiring low "fan-in" and "fan-out" (e.g., parametric phase-locked oscillators and tunnel diodes) may prove feasible.

Modulo-r Carry Sequences

The modulo-r independent–dependent carry adder has a special advantage for radices greater than 2, due to the frequency of occurrence of the dependent carries varying inversely as the radix. Consider equations 3(A) and 3(B). Note that the probability of the occurrence of a dependent carry equals the probability of no overlap of the lines of $+1$'s or -1's. Hence this probability is

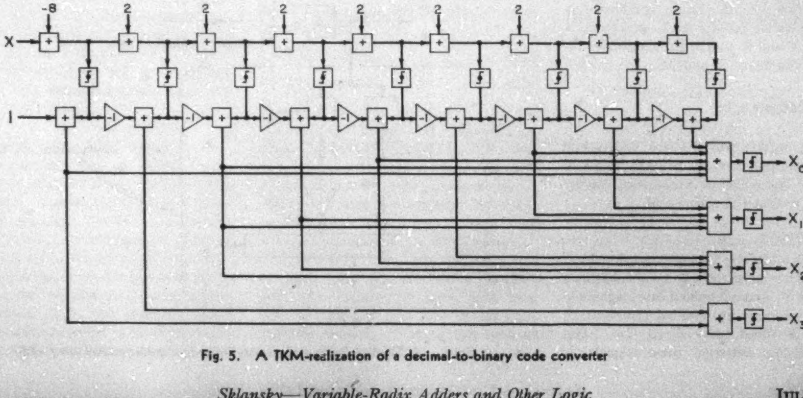


Fig. 5. A TKM-realization of a decimal-to-binary code converter

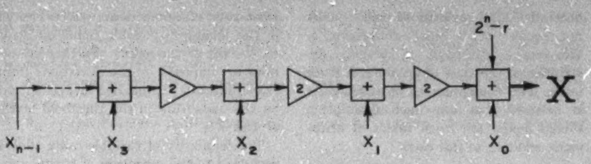


Fig. 6. A TKM-realization of a binary to r-ary converter. The value of n is chosen to be the smallest integer satisfying the inequality $2^n \geq r$

$1/r$. As a result one would expect substantial reductions in the lengths of the carry sequences when the radix is increased. This effect is enhanced by the fact that for a given value of adder capacity, r^n, the summand length, n, is reduced when r is increased, which should produce the following interesting phenomenon: an increase in the average computation speed of the adder when the radix is increased. In the TKM implementation, this would be obtained without increasing the amount of hardware.

This phenomenon is further substantiated by an approximate evaluation of the average length of the longest sequence of 0-carries or 1-carries. The result, obtained by an extension of a formula derived by Burks, Von Neumann, and and Goldstine,[5] is

$$a_n(r) \geq \bar{a}_n(r) = \log_r [2 \log_2 (r^n)] \qquad (12)$$

where r_n is the capacity of the adder. (This formula is derived in the Appendix.) The right-hand member of equation 12 is an upper bound, $\bar{a}_n(r)$, on the average longest 1- or 0- carry sequence, $a_n(r)$. This expression, which is an approximation to $a_n(r)$ as well as an upper bound, is plotted in Fig. 4 as a function of the radix, r, for various specific values of the adder capacity, r^n. Note that for any particular increase in radix, the reduction in $\bar{a}_n(r)$ is greater for larger values of adder capacity.

It is also noted that increasing the radix from 2 to 5 gives, in all cases, a greater reduction in $\bar{a}_n(r)$ than can possibly be obtained by increasing the radix from 5 to any arbitrarily large number. Furthermore, a radix of 5 results in average longest sequences less than 3.05 digits long even when the adder capacity is as large as 10^{30}. Thus the quinary number system seems potentially advantageous for TKM independent–dependent carry adders.

CODE CONVERTERS

The system of TKM logic can realize r-ary to binary and q-ary to r-ary code converters, where q and r are arbitrary integers greater than 2. Fig. 5 shows an r-ary to binary converter for the case $r=10$; the general case of r-ary to binary conversion is easily deduced from this. In the figure, X represents a 10-valued SAC coded input. The outputs, X_i, are the SAC representations of the binary-coded digits representing x. Thus: if $x=3$, then $X=-3$. The binary coding of x is $(x_3, x_2, x_1, x_0)=(0011)$. Hence $(X_3, X_2, X_1, X_0)=(-1, -1, 1, 1)$.

Fig. 6 shows a TKM realization of a binary to r-ary converter, where the number of binary digits, n, is the smallest integer satisfying $2^n \geq r$. The reason for inserting 2^n-r at the last stage is seen from the following:

$$X = 2\left(\sum_{i=0}^{n-1} x_i 2^i\right) - r + 1 \qquad (13)$$

By equation 9(A),

$$x_i = \frac{1}{2}(X_i + 1) \qquad (14)$$

since the r corresponding to the X_i's is 2. Substituting equation 14 in 13 and sum-

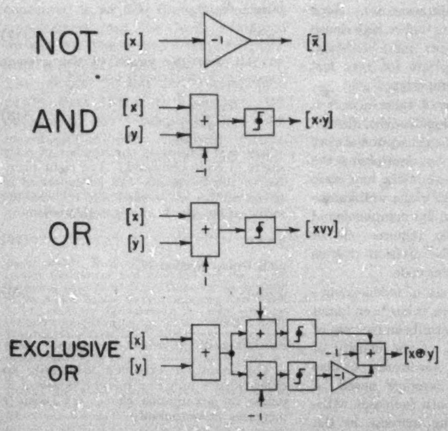
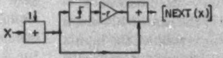
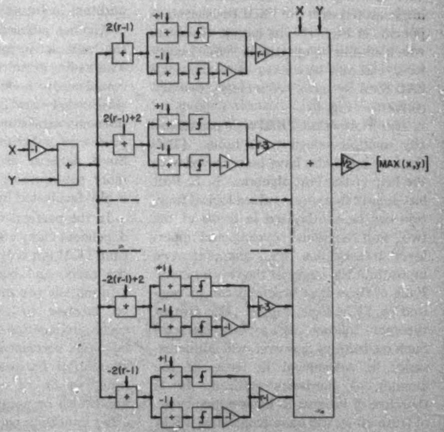


Fig. 7 (above). TKM-realizations of NOT, AND, OR, and EXCLUSIVE OR



(A) CYCLING GATE



(B) UPPER-LEVEL TRANSMISSION GATE

Fig. 8 (right). TKM-realizations of the two Post functions

ming the resulting geometric progression obtains

$$X=\left[\sum_{i=0}^{n-1} X_i 2^i\right]+(2^n-r) \tag{15}$$

Hence 2^n-r is a "correction term" to be added at the last stage. Note that this term is zero when $\log_2 r$ is an integer.

To obtain a q-ary to r-ary converter, one may cascade a q-ary to binary converter and a binary to r-ary converter. More direct methods are perhaps possible; however, the cascade method may be still competitive, because the binary to r-ary converter has a relatively simple structure.

Two-Valued Logic

The system of TKM logic can realize the basic Boolean functions of AND and NOT, and, as a result, any 2-valued truth table. (The reason for this is that the compositions AND and NOT have been shown to form a functionally complete algebra.) Fig. 7 shows the TKM realizations of NOT, AND, OR, and EXCLUSIVE OR.

Many-Valued Logic

The following can now be proved.

Theorem: Any multiple-valued truth table can be realized by a TKM network.

Proof: First note that the input and output variables of a multiple-valued truth table can be coded in the binary system. (The binary-coded decimal system is an example in the case of 10-valued variables.) The conversion of the inputs to binary form can be implemented with an r-ary to binary converter (Fig. 5). The resulting binary truth table can be implemented with the TKM Boolean components of Fig. 7. The output variables, which at this stage are in SAC binary form, can now be converted to the r-ary SAC form by means of a binary to r-ary converter (Fig. 6).

This proves that TKM logic can realize any multiple-valued truth table. (This theorem could also have been proved with the help of the Post algebra. E. L. Post has shown that any r-valued logical function can be synthesized in terms of the two "Post functions": cycling, and upper-level transmission, i.e., mapping two inputs into the larger of the two inputs.[6,7] Each of these Post functions can be realized by TKM logic, Fig. 8. The present theorem follows as a consequence.) Such realizations, however, will not necessarily be economical in terms of the amount of hardware utilized. It is therefore of interest to know which types of truth tables will have economical TKM realizations. (It is expected that economical TKM realizations will usually not require r-ary to binary conversion, or vice versa.) Although a precise description of the class of economical truth tables does not appear to be available, it is believed that any such description should assign the truth tables of arithmetic addition to this class.

Conclusions

It has been shown that the exclusive use of 3-state threshold devices, Kirchoff adders, and linear amplifiers imposes no practical limitation, other than efficiency, on the kinds of truth tables that can be realized. The admissible truth tables may have any finite alphabet of truth values, and the distribution of truth values in the tables may be arbitrary.

With regard to the "3-state" T's, it is noted that in most cases (e.g., in Figs. 2, 3, 5, and 7) a T is called upon to produce only the two extreme values, -1 and $+1$, in which case it can be realized with only one 2-state device.

The full adder and the independent-dependent carry adder seem to be particularly economical in terms of TKM components. Other TKM-realized adders are also expected to prove economical. These adders have the further advantage that the same basic structure may be used for a variety of radices, a change of radix requiring only an adjustment in the gain of one or more amplifiers (or, equivalently, a change in the resistances of one or more attenuation resistors). Since the TKM r-ary adders realize high densities of information per unit hardware, they may be appropriate for very fast addition in future computers.

Another advantage of these adders is the symmetric analog coding (SAC). This coding permits the generation of $r-1$ complements merely by multiplying the SAC-coded signal by -1. The $r-1$ complements can be used for the well-known process of subtraction by complements.[8] Since division usually requires one or more subtractions, the division process is also facilitated by this code.

In the particular case of independent-dependent carry adders, it has been found that TKM not only permits an increase of the radix, and hence of the density of information per unit hardware, but also an increase in the average speed of carry propagation—both increases without any corresponding increase in the amount of hardware. The size of the radix would, of course, be limited by bandwidth or noise considerations. Another limitation occurs in significant digit arithmetic, where the choice of the significant value of a numerical result is limited in the minimum possible round-off error by the size of the radix. Further limitations are imposed by the deviations from exact linearity in the K's and M's, and by the variability in the threshold levels of the T's.

The feasibility of variable-radix adders may lead to the operation of entire computers with variable-radix hardware, thereby making adaptive radix selection possible at the hardware level.

It is hoped that one service of this paper will be to indicate the potential utility of using combinations of analog and threshold devices for digital functions that were previously obtained with strictly digital devices.

Appendix. Derivation of the Formula for an Upper Bound on the Longest Sequence of 0-Carries or 1-Carries in r-ary 2-Summand Addition

The upper bound, and approximation given by equation 12 are derived here.

The starting point is the upper bound derived in the appendix of reference 3 for the average longest sequence of 0-carries or 1-carries in 2-summand *binary* addition:

$$a_q(2)\leq\log_2(2q) \tag{16}$$

where q is the length of either summand, in *bits*. This was obtained by extending a result of Burks, Goldstine, and Von Neumann.[5]

Now consider an r-ary addition, where r is an integral power of 2, i.e.,

$$r=2^p \tag{17}$$

In this case the length of the average longest carry sequence is bounded by

$$a_n(r)\leq\frac{\log_2(2q)}{p} \tag{18}$$

where n is defined as the number of r-ary digits in each summand. The right member of this inequality can be expressed in terms of the adder capacity, r^n, and the radix, r, by means of the substitutions

$$q=np \tag{19}$$

and, from equation 17,

$$p=\log_2 r \tag{20}$$

As a result,

$$a_n(r)\leq\log_r[2\log_2(r^n)] \tag{21}$$

This result is extended arbitrarily to cases where r is not an integral power of 2, under the assumption that $a_n(r)$ versus r increases monotonically.

References

1. ANALOG METHODS (book), W. J. Karplus, W. W. Soroka. McGraw-Hill Book Company, Inc., New York, N. Y., 1959, chap. 4.

2. SIGNIFICANT DIGIT COMPUTER ARITHMETIC, N. Metropolis, R. L. Ashenhurst. *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., vol. EC-7, no. 4, Dec. 1958, pp. 265–67.

3. AN EVALUATION OF SEVERAL TWO-SUMMAND BINARY ADDERS, J. Sklansky. *Ibid.*, vol. EC-9, June 1960, pp. 213–26.

4. FAST CARRY LOGIC FOR DIGITAL COMPUTERS, B. Gilchrist, J. H. Pomerene, S. Y. Wong. *Ibid.*, vol. EC-4, Dec. 1955, pp. 133–36.

5. PRELIMINARY DISCUSSION OF THE LOGICAL DESIGN OF AN ELECTRONIC COMPUTING INSTRUMENT, A. W. Burks, H. H. Goldstine, J. Von Neumann. The Institute for Advanced Study, Princeton, N. J., June 28, 1946.

6. SYNTHESIS OF N-VALUED SWITCHING CIRCUITS, R. D. Berlin. *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, vol. EC-7, no. 1, Mar. 1958, pp. 52–56.

7. INTRODUCTION TO A GENERAL THEORY OF ELEMENTARY PROPOSITIONS, E. L. Post. *American Journal of Mathematics*, Baltimore, Md., vol. 43, 1921, pp. 163–85.

8. ARITHMETIC OPERATIONS IN DIGITAL COMPUTERS (book), R. K. Richards. D. Van Nostrand Company, Inc., Princeton, N. J., 1955.

Ultrahigh-Temperature Electronic Transformers

I—Design Considerations

J. F. RIPPIN
NONMEMBER AIEE

H. B. HARMS
ASSOCIATE MEMBER AIEE

GLENN WALTERS
NONMEMBER AIEE

OVER THE PAST 4 years, military groups have had under development components and accessory parts for electronic equipment destined for ultrahigh-temperature and intense nuclear radiation applications. High-speed flight, an approach to miniaturization and reentry problems all promise ambient temperatures for electronic gear as high as 500 C (degrees centigrade). The anticipated launching of a nuclear-powered aircraft introduced the new environmental requirement, nuclear radiation. One of the component developments concerned is for 500 C electronic power transformers and has been sponsored by the Air Force under two contracts, *AF33(616)-3623* and *AF33(616)-5579*. The purpose of this and a companion paper, Part II,[1] is to summarize the outcome of this development in a form meaningful to design engineers. Design considerations, the subject of this paper, examines the basis for the selection of materials, their mechanical arrangement, and the calculation of coil temperature rise. Using these essential data, Part II[1] then presents a digital computer optimization procedure that yields a minimum-weight transformer design.

Development Objectives

Development activity was concentrated to the extent that the transformer electrical ratings were selected for 400-cps (cycles per second) frequency and a maximum of 1,500 volts rms and 1-kva (kilovolt-ampere) output. The goal for the operating life was 1,000 hours. The environmental objectives included the major ones of 500 C temperature, intense nuclear radiation up to 10^{12} fast neutrons and gamma photons/cm^2/sec. (per square centimeter per second), and air density corresponding to 100,000 feet altitude. The mechanical objectives included 50 *g* shock, 15 *g* vibration up to 2,000 cps, and resistance-to-thermal cycling and moisture. These were to be withstood separately or in such combinations as high ambient temperature with nuclear radiation, shock, and vibration.

Development Approach

The basic method of the development work was to use tests and studies of individual materials for preliminary screening and accumulation of performance data. Co-ordinating these with a compatible mechanical arrangement, complete units were then built from the most promising individual materials and retested under conditions of high temperature, nuclear radiation, and other environments. This limited the more complicated life and environmental tests to a relatively small number of systems or combinations of materials, while giving a reasonably sound basis for selecting these combinations. Selection of a system led finally to the temprature rise and optimum design investigations.

Work was directed toward eventual construction of transformers and inductors using layer-wound coils assembled on the center leg of an *E-I*-lamination core configuration. This goal required investigating magnetic materials for the core, the conductor materials, and the insulation for the wire, as well as spool materials for the winding forms, sheet insulation to separate the layers of the coils, potting materials and impregnants, terminal design and spacing, protective schemes, and materials and processes for making electrical joints. A view of this general design concept is shown in Fig. 1.

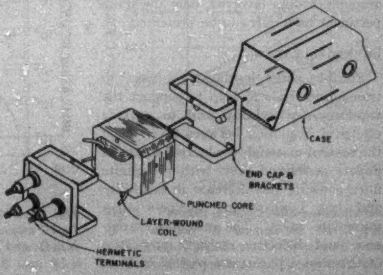


Fig. 1. Hermetic construction concept for 500 C electronic power transformer development

---

Paper 60-231, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. J., January 31–February 5, 1960, and represented for discussion only at the Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 30, 1959; made available for printing November 18, 1960.

J. F. RIPPIN is with the United States Air Force, Gentile Air Force Depot, Dayton, Ohio; H. B. HARMS and GLENN WALTERS are with the General Electric Company, Fort Wayne, Ind.

This paper is an extension of CP58-208 presented at the 1958 AIEE Winter General Meeting. The many contributions of members of the Specialty Transformer Department, the Fort Wayne (Ind.) Laboratory, and the General Engineering Laboratory, all of General Electric Company, to new knowledge summarized in this paper are gratefully acknowledged.

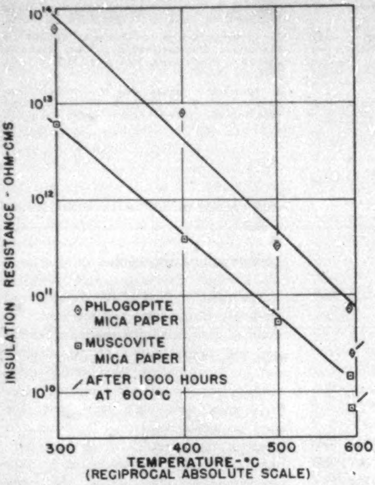
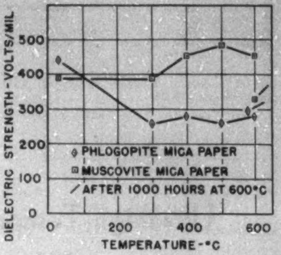


Fig. 2 (left). Insulation resistance temperature–time data for 0.004-inch-thick mica-paper sheet insulations

Fig. 3 (right). Dielectric strength temperature–time data for 0.004-inch-thick mica-paper sheet insulations



Materials Test Results

The primary independent variable considered in this section is temperature. Extremely high temperature makes many materials completely useless. Organic materials as they are known today are completely deteriorated by a 500 C temperature. Therefore, one must rely on the inorganic compositions, some of which may also be found unusable. Even some metals fail to perform their intended function acceptably. The nickel steels for the most part are nonmagnetic at 500 C, and pure copper experiences severe oxidation above 300 C. All metals, including stainless steels, undergo a large reduction in mechanical strength. A thorough screening and evaluation of potentially useful materials had to be made to determine the full impact of temperature effects on transformer design practices. Because of the large amount of data involved, only the more pertinent materials can be presented.

Sheet Insulations

Flexible sheet insulation is used in transformer coils to help to support the windings mechanically and to provide electrical insulation between layers. Of the several materials evaluated[2] the mica papers performed best, with phlogopite mica showing less decrease in flexibility and less increase in thickness than muscovite mica. The insulation resistance and dielectric strength properties of these mica papers are plotted in Figs. 2 and 3 as a function of temperature. Based on a sampling of ten specimens, the insulation resistance was observed to decrease by a factor of 10^3 over the temperature range 300–600 C. A 1,000-hour aging period had a small adverse effect. Drops in dielectric strength of about 25% were attributable to either temperature or aging.

Wire Insulations

A wire insulation should not only be able to withstand the mechanical stresses of winding, but should also have consistent and reasonably good dielectric strength and occupy as little space as possible. Wire insulation and conductor materials were evaluated in simple specimen form on a 1/2-inch-diameter stainless-steel mandrel.[3] Glass-served silver wire had the best all-around performance. Anodized aluminum displayed outstanding electrical properties and thermal stability and had the added advantage of a very thin film build. However, processing and application problems with the base metal require solution before it will become useful at 500 C. Insulation resistance and dielectric strength data on these two insulations are shown in Figs. 4 and 5. Insulation resistance decreased by approximately a 10^4 factor and was not significantly affected by aging time. Dielectric strength decreased for both materials up to 300 C and remained steady thereafter. The large drop of the glass-served material can be attributed to the complete removal of the silicone varnish by the high temperatures. The glass-served insulation on nickel-clad copper wire gave very similar results.[3]

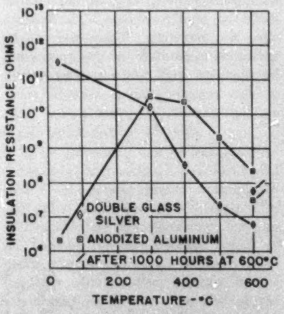


Fig. 4. Insulation resistance temperature–time data for double-glass silver (4-mil build) and anodized aluminum (0.4-mil build) wire

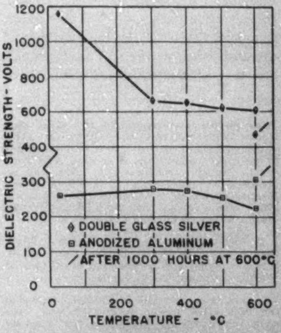


Fig. 5. Dielectric strength temperature–time data for double-glass silver (4-mil build) and anodized aluminum (0.4-mil build) wire

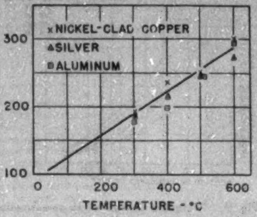


Fig. 6. Per-cent change of conductor electrical resistance as a function of temperature

Conductor Materials

Such considerations as temperature rise and regulation require that transformer conductors have a low-resistance as well as a low-temperature coefficient of resistance. Materials normally selected for their good conductivity, such as silver, aluminum, and copper, must also be considered for 500 C applications. Copper is frequently protected with a nickel cladding where temperatures above 250 C are involved. Fig. 6 shows the effect of temperature on the resistance of these materials. The data indicate an average increase for all three materials of 290% over the temperature range 30–600 C. Coil losses and coil voltage drops are thus altered by a factor of nearly 3 to 1.

The conductor suited best to the requirements of this development was fine silver, because of its good conductivity, corrosion stability, and ease of joining to other metals. Nickel-clad copper is a strong contender but it undergoes a permanent resistance increase of about 35%, caused by diffusion of the nickel into the copper, when exposed to 600 C for 1,000 hours.[3]

Magnetic Materials

For electronic power transformers the magnetic material characteristics of greatest interest are core loss and excitation. Initial studies,[4] including such factors as Curie temperature, processing refinement, and availability, turned specific attention to a $3^1/_4$% silicon grain-oriented steel. It was evaluated in 0.010-inch-thick punched *UI* and *EI* configurations.

Test data[4] are summarized in Figs. 7 and 8 for core loss and excitation as a function of temperature. Because of a resistivity increase in the iron, the core loss decreases approximately 35% from room temperature to 550 C. The data indicate that the phenomenon is independent of flux density. Excitation-temperature phenomena, on the other hand, are dependent upon flux density. Below 50–60 kilolines/inch², the temperature effect is favorable; above that range it becomes gradually more adverse. At 90 kilolines/inch² and 550 C, for example, there is approximately a 70% increase in excitation over that required at 30 C. The over-all performance of the grain-oriented silicon steel is generally acceptable, although some reduction from low-temperature practice may be necessary in operating flux density.

Impregnants, Filler Materials

Protective schemes for the core and coil centered around a metallic enclosure augmented by a filler material to occupy any air spaces within, and by an impregnant to give strength to the coil. The mechanical requirements demand that relative movement of parts within the enclosure be negligible, while the high temperature requires that this be accomplished through the use of inorganic compositions which for the most part are in the form of vitreous ceramics, refractory cements, and inorganic solutions. Extended efforts[5,6] to apply as an alternate enclosing technique a vitreous ceramic as an encapsulant failed because of the brittleness of the ceramic. This material, therefore, has limited use in transformer protective systems. However, a satisfactory refractory cement, consisting of a mixture of magnesium oxide, aluminum oxide, and aluminum phosphate solution,[6] was devised for the filler material. Its brick-like structure gives rigid support to the core and coil. Trials with aluminum phosphate alone suggested its serviceability as a coil impregnant when applied under vacuum-pressure cycling.[5] It imparts strength and rigidity to the coil but does not enhance its dielectric properties.

Cements of the basic alumina, magnesia composition typically display insulation resistance values[6] ranging from 10^6 megohm-cm at 200 C to 10 megohm-cm at 500 C. Aging for 1,000 hours at 500 C has no deleterious effects. A check on the modifying effects of aluminum phosphate solution on the coil insulation system was done with measurements on coils and on simple sheet insulation specimens. Results of tests on the sheet specimens are summarized in Table I and suggest that the impregnant neither impairs nor improves the coil dielectric properties.[8]

Air Dielectric

The spacings of terminals are affected primarily by the properties of air. Air becomes a very critical dielectric when its density is decreased by 500 C temperature and a 100,000-foot altitude and its particles are ionized by radiation. Radiation effects are considered in a subsequent section. The effect of altitude alone is shown in Fig. 9, where the voltage breakdown levels of four typical ceramic-bodied hermetic terminals at room temperature are plotted as a function of altitude up to 150,000 feet.[7]

Since the major effect that increased temperature has on air is to decrease its density, the voltage breakdown level for air between electrodes can be obtained directly from altitude chamber data by converting increased temperature to additional altitude. Changing the temperature from 25 C to 500 C is equivalent to an increase in altitude of approximately 22,000 feet at 100,000 feet. For the terminals considered in Fig 9, the voltage breakdown strength decreases approximately exponentially from rated values ranging from 5–17 kv at sea level and room temperature to a 600–1,200-volt range at 122,000 feet and room

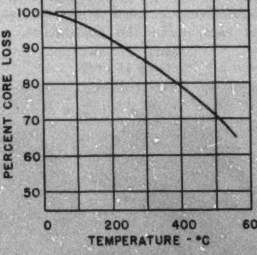


Fig. 7. Per-cent change of 400-cps core loss as a function of temperature for $3^1/_4$% grain-oriented silicon steel. Specimens were punched laminations 0.010 inch thick

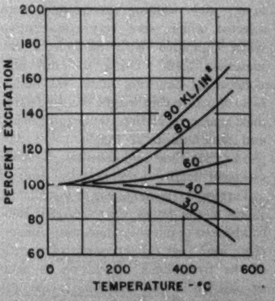


Fig. 8. Per-cent change of 400-cps excitation as a function of temperature for $3^1/_4$% grain-oriented silicon steel. Specimens were punched laminations 0.010 inch thick

Table I. Dielectric Properties of Aluminum-Phosphate Impregnated Phlogopite Mica Paper[5,6]

Property	300 C		600 C	
	Unimpregnated[4]	Impregnated[4]	Unimpregnated[6]	Impregnated[5]
Dielectric strength,* volts/mil	260	258	280	240
Insulation resistance, megohm-cm†	6×10^7	9.3×10^5	7.3×10^4	3.1×10^6

* Averages of at least five specimens.

† Averages of at least ten specimens.

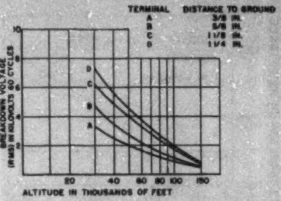


Fig. 9. Dielectric strength of air as a function of altitude for various ceramic-bodied terminals mounted on a flat plate. Data were taken at 30 C. Sea-level ratings are 5 kv (A), 8 kv (B), 16.5 kv (C), 17 kv (D)

temperature (or 100,000 feet and 500 C). Data taken by use of other electrode configurations showed that the hermetic terminal has breakdown strengths that are not as good as those obtained with sphere electrodes, and that above 150,000 feet configuration and spacing are no longer significant variables.[7]

Systems Test Results

With the development objectives directed principally toward a layer-wound coil mounted on punched laminations and with no new data emerging to justify consideration of another configuration, the primary selection of materials from the materials tests and from the filler material and impregnation process developments were then applied to filament and plate transformers typically rated as follows:

1. Filament transformer. 115/15.6 volts, 400 cps, 2.2-ampere load; 15% no load to full regulation, ±10% −65 to 500 C regulation.
2. Plate transformers. 115/700 (mid-tapped) volts, 400 cps, 0.50-ampere direct current into an inductively terminated full-wave rectifier.

Operating steady-state voltage stresses in the coil for turn to turn, layer to layer, and winding to ground were selected which were very conservative when compared with data from the materials tests.[5] Using the full economical *EI* punching shape, the design ratio of coil-to-core weight was a compromise among regulation, temperature rise, and minimum total weight. The Appendix lists these and other of the more pertinent details of the electrical design.

Turning to the mounting or mechanical aspect of the design, several criteria were established to direct its outcome. Among them were that (1) the center of gravity of the transformer should be as close as possible to the mounting plane, (2) the construction should be simple for ease of assembly and low cost, and (3) the design should be capable of accommodating a range of core-stack heights for flexibility of electrical design. Consideration of several metallic structural materials that might be used for this application revealed type *316* stainless steel to be superior to other *300* series and to the *400* series stainless steels.[8] Based upon equal bending and tensile strengths at 1,000 F (degrees Fahrenheit), the type *316* is two to three times lighter and has about the same cost as type *430*.[8] Low-carbon steel cannot compete in this application because of its low strength at 500 C. With a 40,000-psi (pounds per square inch) rupture strength at 1,000 F and applying a safety factor to take into account cyclic heating stresses and other lesser variations, the maximum design stress for type *316* stainless steel was somewhat arbitrarily set at 10,000 psi.

The arrangement of the mechanical parts as they finally evolved is shown schematically in Fig. 1 and by photograph in Fig. 10. The core and coil are fitted into a sleeve-type case which has the mounting feet integral with it. The fit is slightly on the interference side with the core closely set adjacent to a series of embossed ribs. Pairs of brackets attach the core to the end caps which in turn provide mounting space for terminals. A seam weld seals each cap to the case. Holes in the sides of the case provide access for pouring the potting material in place. Covers, silver brazed to seal the holes, and alumina-bodied terminals, silver brazed to the end caps, make this construction a hermetic one. Finally, it was found necessary to add re-enforcing pieces to the mounting feet.

This construction concept has proved to be rugged and adequate. Of nine units vibrated and shocked at the levels specified by the developmental goals, no mechanical failures were noted. Table II lists electrical data taken with a filament transformer as a function of temperature and time. The data indicate stable performance throughout the 1,000-hour aging time (500 hours each at 200 C and 500 C) and little deterioration of the magnetic properties. Resistance of the insulation system decreased with respect to temperature, was unaffected by aging, and fully recovered to its initial high values at 30 C. Electrical failures were observed to occur when daily temperature cycling from 500 C to 30 C was introduced into the life-test cycle. They usually occurred as conductor breaks after 500 hours' aging time in the filament transformers and took place adjacent to the conductor-lead joint in the primary winding.[5] Plate transformer failures on the other hand could be traced to layer-to-layer short-circuited conductors in the high-voltage secondary winding,[5] suggesting the need to rate the sheet insulation even more conservatively.

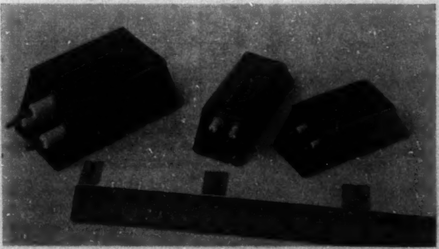


Fig. 10. Hermetic plate (1) and filament (3) transformers and reactor (2) for 500 C nuclear radiation applications

Table II. Electrical Performance of 500 C Filament Transformer[4]

Characteristics	30 C, 25 Hours	200 C, 25 Hours	500 C,* 500 Hours	500 C, 1,000 Hours	30 C, 1,000 Hours
Primary volts	115	115	115	115	115
Primary full-load amperes	0.607	0.625	0.77	0.755	0.623
Primary full-load watts	44.4	43.4	45.4	45.4	43.8
Secondary full-load volts	16.1	15.6		15.2	16.0
Secondary full-load amperes	2.20	2.20	2.20	2.20	2.20
Primary no-load amperes	0.475	0.545	0.73	0.77	0.522
Primary no-load watts	6.8	7.2	9.9	10.2	7.1
Insulation resistance, megohms					
Primary to core	5 000	infinity	9.5	9.0	10^6
Secondary to core	4 700	infinity	7.0	6.5	10^6
Primary to secondary	670	infinity	17.0	15.0	10^6
Winding resistance, ohms					
Primary	4.76	7.04†		11.46	4.88
Secondary	0.1385	0.237†		0.379	0.166

* Life test was conducted at 200 C for 500 hours and at 500 C for 500 hours.
† After 500 hours at 200 C.

Materials and Systems Test Results With Combined High-Temperature and Nuclear-Radiation Environments

The specification of a nuclear radiation environment prompted the performance of a comprehensive series of tests as part of the development program on both transformer materials and systems at the Brookhaven National Laboratory.[5,6,8] Because of the peculiar aspects of conducting tests inside a nuclear reactor there were many special features in the test setup and instrumentation. It is suggested that the references be consulted for a description of these. Flux levels involved in the hole location used in the test reactor at Brookhaven were approximately 1×10^{11} fast neutrons/cm²/sec, 2×10^{12} thermal neutrons/cm²/sec and 1×10^{12} gamma photons/cm²/sec. Results of tests made at these levels can be summarized as follows:

1. A large number of potentially useful materials were tested individually on specially prepared specimens under combined 500 C and radiation environments. Among them were double glass-served wire insulation, mica paper, and grain-oriented silicon steel. The insulation materials for the most part maintained insulation-resistance readings in the 1–100-megohm range with order of magnitude increases occurring consistently at reactor shutdown periods. Data on the magnetic properties of the silicon steel did not show significant changes either as a consequence of the radiation alone or of long-time exposure to it.

2. Radiation adversely affects the dielectric strength of air as evidenced by tests made on cylindrically shaped electrodes spaced at 1/16, 1/18, and 1/4 inch. Air density was also a major variable. Results with the 1/16-inch gap shown in Fig. 11[5,6] indicate a reduction in dielectric strength of about 12% as a consequence of the intense radiation. Reductions as high as 20% were recorded at other temperatures and gap spacings. The low air density in the tests corresponded to a 100,000-foot altitude. Percentage-wise, density did not seem to be a significant variable with respect to the radiation environment.

3. Since the materials that performed best with the 500 C environment also showed up well during exposure to nuclear radiation, they were selected for the systems tests conducted on filament and plate transformers identified in the preceding section. It was intended to operate them in 200 C and 500 C ambients. However, the generation of heat by gamma photons caused the lowest temperature achievable to be 400 C. Comparison of performance data with units that were not irradiated thus had to be done with this inconsistency as shown in Table III. The comparison suggests that there are no gross radiation effects. In all, four units survived 1,300 hours of continuous operation in the combined high-temperature and radiation environments. These circumstances are further illustrated in Fig. 12. An occurrence peculiar to the presence of radiation that should be of considerable interest to designers is the contribution to temperature rise by gamma heating. Extra rises ranging from 20 C to 50 C were observed with the average occurring at 43 C for the filament transformers.[6]

Heat Transfer, Theoretical and Experimental Results

A firm selection of core and coil materials, their configuration, and the mechanical protective scheme, leads, finally, to the consideration of a method for predicting coil temperature rise, an essential aid to the designer of even 500 C transformers.

A prediction method usually is an analytical expression that has a sound theoretical base and that has been correlated with experimental data. Of the many analytical approaches to the problem the one that seems to be the most versatile is the lumped-parameter electric-circuit analogy. The simplest thermal circuit representation of an electronic transformer is shown in Fig. 13. A solution of the coil temperature rise, θ_c, above ambient temperature by the use of Kirchoff's Laws for this circuit yields the expression

$$\theta_c=R_4q_t+(R_1q_c+R_2q_t)\left[\frac{R_3}{R_1+R_2+R_3}\right] \qquad (1)$$

where

θ_c = average rise above ambient temperature of coil, C
R_1 = thermal resistance between coil and core, C/watt
R_2 = thermal resistance between core and case, C/watt
R_3 = thermal resistance between coil and case, C/watt
R_4 = thermal resistance between case and ambient, C/watt
q_c = heat generated by I^2R losses in the coil, watts
q_i = heat generated by losses in the core, watts
q_t = sum of q_c and q_i

The coil loss, q_c, is a function of temperature and is additionally defined by

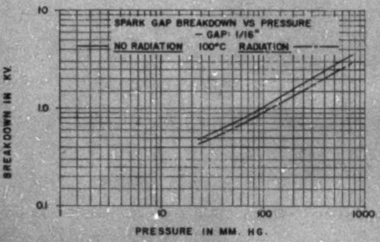


Fig. 11 (left). Dielectric strength of 1/16-inch air gap as function of nuclear radiation intensity and air density

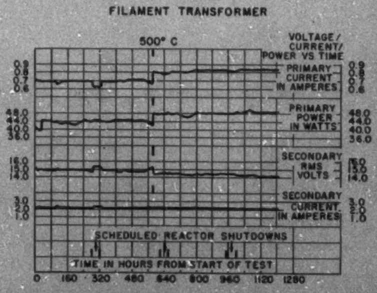


Fig. 12 (right). Electrical performance data of filament transformer during exposure to nuclear radiation

Table III. Comparison of Transformer Data With and Without Radiation, Filament Tranformer

Characteristics	Without Radiation 200 C	Without Radiation 500 C	With Radiation 400 C	With Radiation 500 C
Primary volts	115	115	115	115
Primary full-load amperes	0.625	0.755	0.670	0.730
Primary full-load watts	43.6	45.4	46.0	45.0
Secondary full-load volts	15.6	15.2	14.8	14.7
Secondary full-load amperes	2.20	2.20	2.19	2.21
Primary no-load amperes	0.540	0.77	0.587	0.691
Primary no-load watts	7.1	10.2	8.1	9.0
Insulation resistance, megohms				
Primary to core	50,000	9.0	2.8	1.7
Secondary to core	50,000	6.5	2.5	1.3
Primary to secondary	100,000	15.0	4.2	2.1
Winding Resistance, ohms				
Primary	7.04	11.46	10.0	11.45
Secondary	0.237	0.379	0.312	0.352

$$q_c = q_{co}[1+\alpha(t_a+\theta_c)] \tag{2}$$

where

q_{co} = coil loss at 0 C
α = temperature coefficient of resistivity, 1/C
t_a = ambient temperature, C

The principal limitations of this representation are (1) that it is assumed that no temperature gradients exist in either the coil or the core, and (2) that the assignment of heat-flow paths for the calculation of thermal resistances is somewhat arbitrary.

The conduction thermal resistances R_1, R_2, and R_3 each represent several paths in series and parallel and, consequently, are defined by unwieldy expressions. Basically, however, the individual path resistance has the form

$$R = \frac{l}{kA} \tag{3}$$

where

l = length of heat flow path parallel to the assumed heat flow direction, inches
k = thermal conductivity of material making up the path, watt inches/inch^2 C
A = area of the heat-flow path perpendicular to the assumed heat-flow direction, inch^2

Figs. 14, 15, and 16 qualitatively depict the heat-flow paths selected for this application. Resistance R_1 for coil-to-core conduction is made up of

R_{11} = heat flow through the spool and an air space to the center leg of the core
R_{12} = heat flow through the outer coil wrapper to the outer legs of the core
R_{13} = heat flow through the coil margins to the cross legs of the core

Realizing that these are all in parallel, R_1 becomes

$$R_1 = \frac{R_{11}R_{12}R_{13}}{R_{12}R_{13}+R_{11}R_{13}+R_{11}R_{12}} \tag{4}$$

Similarly, resistance R_2 for core-to-case conduction is made up of

R_{21} = heat flow parallel to the laminations through a small air space to the case
R_{22} = heat flow through the filler material to the case
R_{23} = heat flow through the filler material to the end caps

R_2 can then be more accurately identified by

$$R_2 = \frac{R_{21}R_{22}R_{23}}{R_{22}R_{23}+R_{21}R_{23}+R_{21}R_{22}} \tag{5}$$

Finally, resistance R_3 for coil-to-case conduction is made up of

R_{31} = heat flow from the outer periphery of the coil, through the filler material and to the end caps
R_{32} = heat flow from the coil margins, through the filler material and to the case

R_3 then becomes

$$R_3 = \frac{R_{31}R_{32}}{R_{31}+R_{32}} \tag{6}$$

The thermal resistance R_4 involves free convection and radiation modes of heat transfer and can be expressed as follows

$$R_4 = \frac{1}{H_t} \tag{7}$$

where

$H_t = H_c + H_r$
H_c = convection heat transfer coefficient, watts/C
H_r = radiation heat transfer coefficient, watts/C

More specifically for this application H_c and H_r[9,10] are

$$H_c = C_1\left(\frac{\theta_s}{T_a}\right)^{1/4} \tag{8}$$

$$H_r = C_2(\theta_s+2T_a)[(\theta_s+T_a)^2+T_a^2] \tag{9}$$

where

$C_1 = (0.0118A_1L_1^{-1/4}+0.0092A_2L_2^{-1/4})\sqrt{P_r}$
$C_2 = 3.33\times10^{-11}(A_1+A_2)$
A_1 = area of top surface of the case, inch^2
$L_1 = \frac{\text{length}\times\text{width}}{\text{length}+\text{width}}$ of the top surface of the case, inches
A_2 = sum of areas of vertical case surfaces, inch^2
L_2 = case height, inches
θ_s = surface temperature rise, C
P_r = ratio of pressure at the given altitude to that at sea level
T_a = absolute ambient temperature, degrees Kelvin

It is assumed in the H_r term that the radiation emissivity factor is 0.9 and that the configuration factor is unity. Because of the nonlinearity of the expressions for H_c and H_r a trial-and-error solution must be resorted to in order

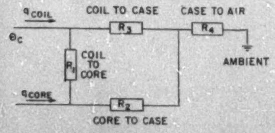


Fig. 13. Lumped-parameter thermal circuit representation of electronic transformers for derivation of expression for coil temperature rise

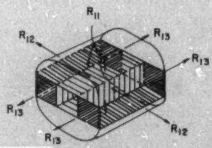


Fig. 14. Heat-flow paths defining conduction resistance, R_1, of Fig. 13

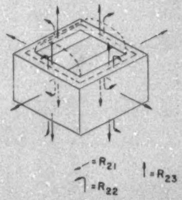


Fig. 15. Heat-flow paths defining conduction resistance, R_2, of Fig. 13

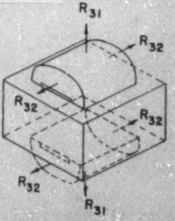


Fig. 16. Heat-flow paths defining conduction resistance, R_3, of Fig. 13

to satisfy the average surface temperature rise, θ_s, as determined by

$$\theta_s=\frac{q_t}{H_c+H_r} \tag{10}$$

With this background it is expedient to reform equation 1 as

$$\theta_c=\theta_s+C_cq_c+C_hq_t \tag{11}$$

where

$$C_c=(R_1+R_2)\left[\frac{R_3}{R_1+R_2+R_3}\right]$$

$$C_h=R_2\left[\frac{R_3}{R_1+R_2+R_3}\right]$$

To correlate equation 11 with experimental data the filament and plate transformers identified in an earlier section and representing a reasonable spread of core dimensions (*EI-50* and *EI-87*) were subjected to temperature rise tests at 30 C, 250 C, and 500 C ambient temperatures. Rises were measured by thermocouples located inside the case and adjacent to the coil. Data obtained with the filament transformer are represented by the dashed curves in Fig. 17. Correlation attempts proceeded along several lines with the one giving best results being centered around modification of the H_t term by introducing a factor F_A into equation 10 (or 7) to yield

$$\theta_s=\frac{q_t}{F_AH_t} \tag{12}$$

F_A, it was found, varied inversely with the core-center leg width, a, as

$$F_A=0.875/a \tag{13}$$

Thus, $F_A=1$ for the plate transformer with the *EI-87* core and $F_A=1.75$ for the filament transformer with the *EI-50* core. In arriving at this result only the 28 C ambient data were considered and the materials making up the internal conduction paths were all assigned a conductivity value of 0.0043 watt inch/inch^2 C.

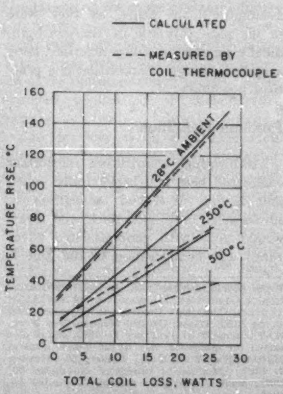


Fig. 17. Measured and calculated temperature rises for 500 C filament transformer as function of coil loss and ambient temperature. Core loss is approximately 5 watts

The solid curves of Fig. 17 indicate the accuracy obtained by this procedure with the filament transformer. It should first be observed that for equal coil losses the temperature rise decreases substantially as the ambient temperature increases. Both the experimental and analytical results confirm the occurrence of improved radiation heat transfer from the transformer surface at higher ambient temperatures. Correlation of the results at 250 C and 500 C, however, were far from acceptable. The large differences can be partially attributed to the difficulty of measuring an ambient temperature inside a furnace beset by gradients and unknown wall temperatures. Since no attempts were made to control closely this aspect of the experiment it was, therefore, not considered pertinent to correlate with doubtful data. With the high-ambient cases, then, the theoretical rise predictions should be given more credence.

Summary

The big step from today's 200 C ambient temperature for electronic equipment to a predicted 500 C is largely a materials and processes problem for most components including, particularly, power transformers. The realization that flexible organic materials must be removed from the insulation picture entirely (except as temporary strengtheners for construction purposes) and substituted for by more brittle inorganic compositions gives a broad definition to the problems to be encountered with ground-, layer-, and wire-insulating materials. Similarly, conductor material for wire and magnetic materials for the core encounter adverse conditions at 500 C which impose limitations on performance and even eliminate some from application altogether. It is well known that mechanical strength of steel structural members suffers large deratings at 500 C. Low-carbon steel, the workable stand-by for normal transformer usage, for example, becomes too bulky to maintain the required strength. In all material categories an intense nuclear radiation specification introduced still other limiting situations.

Thus it is apparent that the several available materials for transformer applications at 500 C had to be evaluated to determine quantitatively the full impact of high temperature and nuclear radiation on their individual properties. The acceptable and best, if any, among them then had to be assembled into transformers, a protective system devised for them, and the combination tested to establish mechanical compatibility and safety-factor adequateness. The paper describes these steps, as well as an additional one concerning temperature rise prediction, which composed, for the most part, the development program carried out for the Air Force for 500 C nuclear-radiation-tolerant electronic power transformers. Still another step, design optimization, is described in the companion paper, Part II.[1]

Selecting layer-wound coils mounted on punched lamination cores for the systems concept gave initial direction to the program. Among the many materials tested which fit this concept, glass serving for the wire and phlogopite mica paper gave best promise for the coil insulation. Silver was accepted for the conductor because of its corrosion stability, good conductivity, and comparative freedom from joining problems. Grain-oriented $3^1/_4$% silicon steel is applicable for the core of power transformers, although lower flux densities may have to be used. A stainless-steel hermetic case, alumina-bodied terminals, refractory cement for filler material, and an inorganic solution for the coil impregnant evolved as successful protective devices for the core and coil.

Radiation testing indicated that those materials that could survive a 500 C temperature could also remain essentially unaffected by a nuclear radiation environment. Radiation does, however, cause lower air breakdown voltages. In coil-temperature rise considerations, higher ambient temperatures cause increased coil losses, but these are offset by improved radiation heat transfer, leaving a possible net decrease in temperature rise. In general, high temperatures have adverse effects on insulating, magnetic, and structural materials. Despite these effects, however, transformers capable of meeting the electrical and environmental goals of the development appear to be practical.

Appendix. Electrical Design Details for 500 C Filament Transformer

Frequency—400 cps
Voltage—115/15.6

Lamination—*EI-50* 0.010-inch grain-oriented $3^1/_4\%$-silicon steel; 0.85-inch stack

Primary winding—196 turns 0.0126-inch-diameter double-glass silver

Secondary winding—28 turns 0.032-inch-diameter double-glass silver

Spool—0.030-inch $AlPO_4$ bonded built-up muscovite mica paper

Layer, ground, interwinding insulation—0.0045-inch phlogopite mica paper

Current density—2,810 amperes/inch² primary 2,740 amps/inch² secondary

Flux density—80 kilolines/inches²

Turn-to-turn stress—0.59 volt

Layer-to-layer stress—18 volts/mil (test)

Winding-to-ground stress—37 volts/mil (test)

Primary no load watts (30 C)—5.5

Primary no load current (30 C)—0.123

Coil loss (30 C ambient)—1.11 watts

Calculated temperature rise—38 C

References

1. Ultrahigh-Temperature Electronic Transformers; II—Design Optimization, **Glenn Walters.** *AIEE Transactions*, see pp. 302–09 of this issue.

2. Properties of Electrical Insulation at Ultrahigh Temperatures, **G. I. Duncan, M. M. Felger.** *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 77, Nov. 1958, pp. 300–05.

3. Magnet Wire for 600 C Temperature Transformers, **W. W. Wareham.** *Ibid.*, pt. I (*Communication and Electronics*), vol. 78, Mar. 1959, pp. 66–70.

4. An Experimental Study of Magnetic Materials for Use in Ultrahigh-Temperature Electronic Transformers, **H. B. Harms.** *Ibid.*, vol. 77, May 1958, pp. 181–84.

5. Ultra High Temperature (500°C) Power Transformers and Inductors, **H. B. Harms, J. C. Fraser.** *WADC TR59-348*, Wright Air Development Center, Dayton, Ohio, July 1959.

6. Ultra High Temperature Miniaturized Power Transformers and Inductor Materials, **H. B. Harms, J. C. Fraser,** *WADC TR57-492*, Wright Air Development Center, May 1958.

7. Terminal Spacings for High-Altitude and Ultrahigh-Temperature Electronic Transformer Applications, **G. I. Duncan, W. A. Rectanus.** *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 78, Mar. 1959, pp. 16–18.

8. Materials for High Temperature Service, **Clauser.** *Materials & Methods*, New York, N. Y., Apr. 1954.

9. The Effects of Nuclear Radiation on the Electrical Strength of Air, **G. I. Duncan, J. C. Fraser, B. Valachovic.** *AIEE Transactions*, pt. I (*Communications and Electronics*), vol. 79, Mar. 1960, pp. 19–26.

10. The Thermal Evaluation of Air-Cooled Electronic Equipment, **W. Robinson, R. H. Zimmerman.** *Air Force Technical Report no. 6579*, U. S. Air Force, Washington, D. C., Sept. 1952, pp. 130–37.

◆

Discussion

L. J. Steinbach (Bell Telephone Laboratories Inc., Whippany, N. Y.): With respect to the paper, there are two points which we feel deserve more explanation.

Excitation-temperature phenomena are discussed briefly and are shown in Fig. 8. As mentioned, the excitation decreases as the temperature is increased when the flux density is below 50 kilolines/inch². However, when the flux density is greater than 50 kilolines the excitation increases when the temperature is increased. We feel that the reason for this should be explained.

The paper also states that alumina-bodied terminals are silver brazed to the end caps. As we know silver-brazing alloys, the joint may be subjected to temperatures up to about 220 C without appreciable loss in strength. At a temperature of 500 C it would appear that the joint strength would be quite low. It is not clear whether mechanical tests were made at 500 C on assembled transformers.

J. F. Rippin, H. B. Harms, and **Glenn Walters:** The question concerning the effect of temperature on excitation of silicon steel received attention in reference 4 of the paper. It was suggested there that, within the saturation limits of the steel, the permeability increases as temperature increases. The lower excitation at an induction of 50 kilolines/inch² and smaller are, therefore, a direct outcome of this. However, superimposed upon this effect is the gradually lowering saturation value which finally at the Curie temperature of 740 C yields a totally nonmagnetic condition. The substantially higher excitation values at inductions greater than 50 kilolines/inch² emphasize this tendency.

It is acknowledged, as the discusser suggests, that the silver-brazing alloy and the terminal-to-end-cap joint of which it becomes a part weaken as temperature increases. No data are available to indicate the degree of this weakening. Vibration and shock tests conducted at 500 C on assembled transformers have shown that the joint was entirely adequate in this particular application. Some help could have come from the support the terminals received internally by the potting compound, but in one construction type tested this was known to be negligible. It is felt, in any case, that a high degree of confidence can be placed in the adequacy of the terminal-to-end-cap joint as used here. Reference 5 of the paper documents these statements.

Ultrahigh-Temperature Electronic Transformers

II—Design Optimization

GLENN WALTERS
NONMEMBER AIEE

A DESIGN OPTIMIZATION technique for 500 C (degrees centigrade) electronic transformers makes possible the determination of core dimensions, flux density, and other parameters needed to achieve minimum weight. Certain constraints on these parameters, however, must be satisfied, chief of which relate to performance specifications: power output, regulation, and temperature rise.

Transformer optimization typifies a general class of problems, to which the same solution technique is applicable, and which are characterized by a set of nonlinear equations that must be resolved before a design will be acceptable. Since the equations are fewer in number than the variables, there are sufficient degrees of freedom to minimize simultaneously such a function as weight, volume, losses, or cost, which is regarded as the criterion of the optimum design. The complexity of the constraint equations and criterion function usually precludes an analytical formula for the optimum design, unless many simplifying and perhaps dubious assumptions are made, as they were in the past. By use of the quasi-gradient computer procedure described here, a realistic design optimization is a practical possibility.

Problem Definition

The ultrahigh-temperature transformer considered here is a hermetically sealed unit having *EI*-shaped laminations of silicon steel, 0.010-inch thick, and layer-wound coils with glass-insulated silver wire. The configuration of the unit is

---

Paper 60-230, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960, and re-presented for discussion only at the Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 30, 1959; made available for printing November 23, 1960.

Glenn Walters is with General Electric Company, Fort Wayne, Ind.

The work described was supported in part by Wright Air Development Center under Contract no. *AF33(616)-5579*.

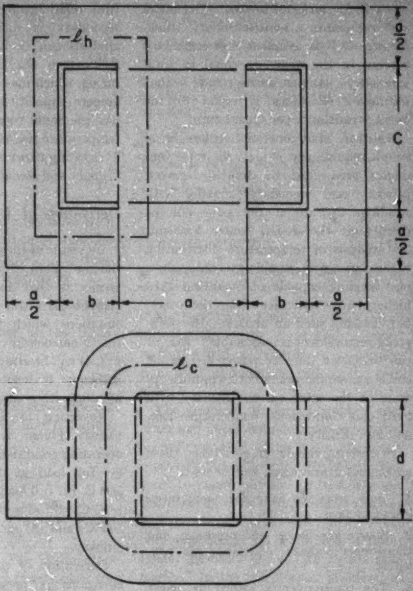
shown in a companion paper,[1] and the configuration of core and coil in Fig. 1.

Four core dimensions, in Fig. 1, are to be determined: center-leg width a, window width b, window length c, and stack height d. Widths of outside legs and yokes are fixed proportions of the center-leg width. Except for variations in those dimensions, the basic mechanical design will remain fixed. Thus, consideration need not be given to specifications that affect mechanical design and choice of insulating materials, such as vibration and shock resistance, thermal cycling, nuclear radiation resistance, and insulation tests. Operation is assumed to be at sea level and at voltages low enough not to affect conductor, lead, and terminal spacings.

The specifications of principal concern are:

1. Power output: up to 1,000 volt-amperes.
2. Regulation, no load to full load at 300 C ambient: 15% maximum.
3. Regulation, full load over the ambient temperature range (−65 C to 500 C): 10% maximum.
4. Temperature rise of coil at 500 C ambient: 75 C maximum.
5. Frequency: 400 cps (cycles per second).
6. Types of loads: tube filaments and full-wave rectifier circuits.

Fig. 1. Configuration of core and coil



Full-load current and the voltage impressed on the tranformer primary winding are considered independent of the ambient temperature. (Values of current and voltage do not have to be specified with the method of analysis given herein.) Limitations on the range of certain design parameters, particularly the core dimensions, exist, and will be described later. Transformer weight was selected as criterion of the optimum design because of the expected application of these transformers.

The problem is to determine core dimensions, number of turns, conductor sizes, flux density, current density, and other parameters that yield a design which will meet specifications and have minimal weight. In mathematical terms, the problem may be expressed by letting $x_1(i=1, \ldots, n)$, denoting the design variables, and finding values of the x_1 that satisfy

$$\alpha_i \leq x_i \leq \beta_i \qquad i=1, \ldots, n \tag{1}$$

$$f_j(x_1, \ldots, x_n)=0 \qquad j=1, \ldots, m<n \tag{2}$$

and

$$g(x_1, \ldots, x_n)=\text{a minimum} \tag{3}$$

The functions f_j represent performance specifications and certain other restrictions imposed on the design; α_i and β_i are constants; and function g is the criterion function. Values that satisfy 1 and 2 constitute an acceptable design. An optimum design may be defined as one that is acceptable and one for which the criterion function is a minimum.

The preceding statement assumes that performance specifications can be expressed as relations 1 and 2. This is possible for transformer optimization, as will be shown, by a proper selection of parameters to be treated as variables, and by introduction of suitable simplifying assumptions.

Assumptions

In optimizing design, much computation can be avoided and adequate accuracy obtained if the coil is treated as a simple continuous structure. Hence, design equations will be developed for a "model" transformer having two isolated windings, a primary and a secondary, each comprising one turn of a single homogeneous conductor. Furthermore, the total conductor cross-sectional area is assumed to be a fixed, known fraction F_c of the window area.

This treatment ignores the fact that wire diameter, number of turns, turns per layer, thickness of layer insulation, and clearances are discrete variables. The principal deficiency concerns the uncertainty of the conductor-area fraction F_c. However, the approximate range of F_c can be determined, and optimum designs calculated for several typical values. The only alternative is a more complicated analysis, using wire diameter, number of turns, etc., as variables.

An advantage, in addition to simplification of the analysis, is that actual values of impressed voltage and load current are not needed, and in fact, are unknowns as related to the model. Thus, a design need not be reoptimized for variations in voltage and current—provided specifications on power output, regulation, and temperature rise do not change. While eliminating the need for actual values of power output is desirable, this is not possible.

The model is related to the actual transformer by the equations

$$V=v/n,\ I=ni,\ J=j,\ R=r/n^2,\ B=b \tag{4}$$

V, I, J, R, B, and $N(=1)$ are voltage, current, current density, resistance, flux density and number of turns for the model; v, i, j, r, b, and n are corresponding parameters for the actual. These equations form the basis for an actual design when V, I, v, and i are known. The model also represents a multisecondary transformer—if it is assumed that secondary current in the model is equal to total secondary ampere-turns, and that secondary resistance is equal to the average resistance per square turn.

Another simplification ensues if a simple requirement is met: the total conductor area must be divided between primary and secondary, making both windings operate at the same current density at the design center ambient. An analysis of temperature distribution within the coil would probably show one winding capable of operating at a higher current density than the other, but, lacking such an analysis, the foregoing requirement is reasonable. Also, it can be shown that the power loss in the coil is a minimum when both windings operate at the same current density with fixed core dimensions, temperature rise, and flux density.

In deriving the design equations, these additional assumptions will be used:

1. Both windings have the same mean length of turn.
2. Source impedance, lead resistance, and transformer leakage inductance are negligible.
3. The transformer is loaded by linear resistance; all currents, voltages, and fluxes are sinusoidal.
4. Under no load, voltage drops and temperature rise can be neglected.
5. The core loss and excitation data for the standard *EI* laminations are applicable to the range of *EI* shapes to be considered.
6. The thermal coefficient of resistivity is a constant over the ambient temperature range.

Some comment on assumption 3 is in order. This assumption is commonly made in designing filament transformers. If the secondary is loaded by a rectifier circuit, the load is nonlinear and, in the case of a full-wave rectifier, the ampere-turn relationship between primary and secondary is different. However, the analysis can be applied to plate transformers on the basis of an equivalent rating, which may be defined as the volt-ampere capacity of a transformer that has the same total conductor area and temperature rise as the plate transformer. This rating turns out to be the average of primary and secondary volt-amperes.

Derivation of Design Equations

By virtue of the preceding assumptions, the equivalent circuit of the model transformer is that shown in Fig. 2. The symbols shown in that figure and in the equations which follow are defined in the "Nomenclature." A subscript k $(=1, 2, 3, 4)$ will be attached to each symbol, as necessary, to denote the value at a particular operating condition.

Reverting to the specifications previously given, it is seen that four operating conditions must be considered: (1) full load at design center ambient 300 C, (2) full load at minimum ambient −65 C, (3) full load at maximum ambient 500 C, and (4) no load at design center ambient.

Derivation of the transformer design equations is presented in two parts. First, the basic voltage and current relations are derived, and formulas for power output, regulation, temperature rise, and transformer weight are given. Second, basic equations are related to performance specifications and to restrictions discussed under "Assumptions"; then constraints similar to relations 1 and 2 are set up.

BASIC EQUATIONS

The load component of primary current is I_{sk}, since primary and secondary windings have one turn each. The component I_{hk} of primary current due to core loss is in phase with the load component and the magnetizing current I_{ek} in quadrature, so the total primary current is given by

$$I_{pk}=[(I_{sk}+I_{hk})^2+I_{ek}^2]^{1/2} \tag{5}$$

Since leakage flux is assumed negligible, the voltages induced in primary and secondary are equal. Hence

$$E_k=E_{pk}=E_{sk}=C_E A_h B_k \tag{6}$$

where

$$C_E=\pi\sqrt{2}f\times 10^{-8} \tag{7}$$

Terminal voltages under full-load conditions are given by

$$V_{pk}^2=[E_k+R_{pk}(I_{sk}+I_{hk})]^2+[R_{pk}I_{ek}]^2 \qquad k=1,2,3 \tag{8}$$

$$V_{sk}=E_k-R_{sk}I_{sk} \qquad k=1,2,3 \tag{9}$$

and for the no-load condition, since voltage drops are assumed negligible, by

$$V_{p4}=E_4=V_{s4}=C_E A_h B_4 \tag{10}$$

The regulation R_n, no load to full load at design center ambient, is defined by

$$R_n=\left|\frac{V_{s4}-V_{s1}}{V_{s1}}\right|=\left|\frac{V_{p4}-V_{s1}}{V_{s1}}\right| \tag{11}$$

and regulation R_f, full load over the ambient temperature range, by

$$R_f=\left|\frac{V_{s2}-V_{s3}}{V_{s2}}\right| \tag{12}$$

The power output of the transformer is simply the product of secondary terminal voltage and load current or, using equation 9,

$$P_{ok}=(E_k-R_{sk}I_{sk})I_{sk} \tag{13}$$

The power loss in the coil is

$$P_{ck}=R_{pk}I_{pk}^2+R_{sk}I_{sk}^2 \tag{14}$$

and the core loss is

$$P_{hk}=G_h W_k \tag{15}$$

The excitation components of primary current are given by

$$I_{hk}=P_{hk}/E_k \tag{16}$$

and

$$I_{ek}=l_h H_k \tag{17}$$

Data for computing core loss per pound W and magnetizing ampere-turns per inch H are given in the figures cited in reference 2. Those figures show curves of W and H versus flux density over the range 30 to 100 kilolines per square inch at 30, 250, 400, and 550 C; also shown are correction factors for lamination size. The curves can be approximated within 2% accuracy by a fifth-order polynomial in B. Hence, W and H at each operating condition can be calculated by

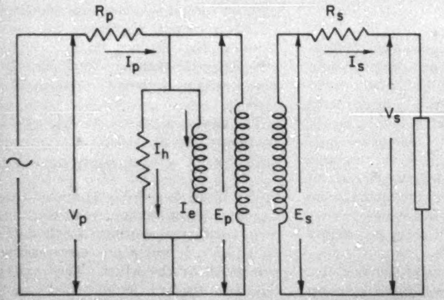


Fig. 2. Equivalent circuit for model transformer

$$W_k = C_{FW} \sum_{i=0}^{3} \left[\sum_{j=0}^{3} C_{Wij} t_{hk}^{j} \right] B_k^{i} \tag{18}$$

$$H_k = C_{FHk} \sum_{i=0}^{3} \left[\sum_{j=0}^{3} C_{Hij} t_{hk}^{j} \right] B_k^{i} \tag{19}$$

The core temperatures t_{hk} can be approximated by the ambient temperatures $t_{ak}(=300, -65, 500, 300)$ with one exception: the −65 C ambient is beyond the range of the data, and extrapolation is unreliable. For that ambient, the best procedure is to use the closest available curve. Hence

$$t_{kh} = 300,\ 30,\ 500,\ 300 \tag{20}$$

The correction factors in equations 18 and 19 are functions of the lamination size and flux density. Assuming center-leg width a is an adequate measure of lamination size, the factors are

$$C_{FW} = 1.27 - 0.24a \tag{21}$$

$$C_{FHk} = 1.312 - 0.274a + (5.81 - 5.27a) B_k \times 10^{-3} \tag{22}$$

These equations are valid for values of center-leg width from 0.37 to 1.50.

The total conductor cross-sectional area A_c is assumed to be a fixed fraction F_c of the window area. Let the area of the secondary conductor be σA_c. Then the area of primary conductor is $(1-\sigma)A_c$. The current densities are given by

$$J_{pk} = I_{pk}/(1-\sigma)A_c,\ J_{sk} = I_{sk}/\sigma A_c \tag{23}$$

and winding resistances by

$$R_{pk} = \rho_k l_c/(1-\sigma)A_c,\ R_{sk} = \rho_k l_c/\sigma A_c \tag{24}$$

The resistivity ρ is subscripted by $_k$ because it varies with the temperature of the coil and hence has a different value at each operating condition. The thermal coefficient of resistivity, α, is assumed to be independent of temperature, so the resistivity is given by

$$\rho_k = \rho_0 [1 + \alpha(t_{ak} + \theta_{ck})] \tag{25}$$

where ρ_0 is the resistivity at 0 C.

Lengths, areas, and weights of core and coil are given by equations 26 through 32; also see Fig. 1.

$$l_h = 2(a+b+c) \tag{26}$$

$$A_h = adF_h \tag{27}$$

$$G_h = \delta_h l_h A_h \tag{28}$$

$$l_c = 2a + 2d + \pi b \tag{29}$$

$$A_c = bcF_c \tag{30}$$

$$G_c = \delta_c l_c A_c \tag{31}$$

Total transformer weight is

$$G_T = G_h + G_c + G_e \tag{32}$$

G_e is the weight of case, core clamp, and filling compound, and is given by

$$\begin{aligned} G_e = {} & 2\delta_s \tau [CD + (C+D)(A+0.125)] + \\ & 2(\delta_s - \delta_f) a \tau [B + 2a + 2b] + \\ & \delta_f [BCD - 2bc(a+2b)] \end{aligned} \tag{33}$$

where

$$A = d + 2b + 1.531$$

$$B = 2b + 1.031$$

$$C = a + C + 0.062$$

$$D = 2a + 2b + 0.062$$

Temperature-rise equations have been developed, and are described in some detail in the companion paper. The temperature rise of the coil is given by

$$\theta_{ck} = \theta_{sk} + C_c P_{ck} + C_h P_{hk} \tag{34}$$

and the temperature rise of the transformer surface by

$$\theta_{sk} = (P_{ck} + P_{hk})/H_{tk} \tag{35}$$

H_{tk} is the surface heat-transfer coefficient resulting from convection and radiation, and is given by

$$H_{tk} = F_A C_1 \left[\frac{\theta_{sk}}{T_{ak}} \right]^{1/4} + F_A C_2 \left[\frac{(\theta_{sk} + T_{ak})^4 - T_{ak}^4}{\theta_{sk}} \right] \tag{36}$$

where $T_{ak} = t_{ak} + 273$. The parameters C_c, C_h, C_1, C_2, and F_A are functions of the the core dimensions (and of other parameters which do not appear in this paper). The surface rise is of no particular interest here; it cannot, however, be algebraically eliminated.

For convenience in optimizing the design, core dimensional ratios p, q, and r are introduced; their definitions are

$$p = b/a,\ q = c/b,\ r = d/a \tag{37}$$

Subsequently, a, p, q, and r will be taken as the basic core dimensional parameters rather than a, b, c, and d.

DESIGN CONSTRAINTS

Setting up constraints in the form of relations 1 and 2 is, to some extent, an arbitrary procedure. Consider, for example, the regulation R_n. Letting R_{ns} denote the maximum allowable regulation, it is required that

$$R_n \leq R_{ns} \tag{38}$$

There are at least four different ways to express this requirement:

1. Regard R_n as one of the basic variables x_i. Then, R_{ns} is the upper bound on that variable, and equation 11 constitutes a constraint equation, which must be satisfied by R_n and other variables.

2. Regard R_n as a function of other variables, the function being defined by equation 11. The relationship in equation 38 can be changed by introducing a slack or dummy variable z_n, resulting in

$$0 \leq z_n,\ R_n + z_n = R_{ns} \tag{39}$$

The bounded variable z_n is then taken as one of the x_i.

3. The slack variable also can be chosen, as shown by Brown,[3] so the constraint equation is

$$R_n + Z_n^2 = R_{ns} \tag{40}$$

The variable Z_n is unbounded.

4. Define the function

$$\begin{aligned} f &= 0, \text{ if } R_n \leq R_{ns} \\ &= R_n - R_{ns}, \text{ otherwise} \end{aligned} \tag{41}$$

It is evident that the relationship in 38 is satisfied if $f = 0$.

Methods 1 and 2 will be used for initial formulation of the constraints. Later, method 4 will be referred to in connection with eliminating redundant variables.

The problem statement implies that lower and upper bounds must be imposed on each variable. This is not absolutely necessary but is desirable for several reasons: to insure physical realizability and the existence of an optimum, and to avoid regions where singularities exist. When only one bound on a variable actually exists, the other one can be chosen arbitrarily. In the transformer problem, all lower bounds must be equal to or greater than zero for physical realizability.

Since full-load current and impressed voltage are to be considered independent of ambient temperature, the subscript k on V_p in the basic equations can be omitted; also

$$I_{s1} = I_{s2} = I_{s3} = I_s \tag{42}$$

$$I_{s4} = 0 \tag{43}$$

A set of parameters, which can be regarded as basic variables, is selected after studying the basic equations. The set deemed appropriate is:

Core parameters: a, p, q, r
Flux density: B_1, B_2, B_3, B_4
Load current: I_s
Secondary conductor area fraction: σ
Coil temperature rise: θ_{c1}, θ_{c2}, θ_{c3}
Surface temperature rise: θ_{s1}, θ_{s2}, θ_{s3}

All other parameters are either constants or explicit functions of the basic variables.

Inclusion of temperature rises for design center and minimum ambient conditions might be questionable, since the rise at maximum ambient will undoubtedly be the limiting value. However, θ_{c1} and θ_{c2} are needed in the computation of winding resistances which are, in turn, required for the regulations and also for power output. As temperature rise at no load will be neither limiting nor necessary in the computation of other parameters, it has been omitted.

The applied voltage V_p might be considered a basic variable rather than the no-load flux density, since the two are proportional, as shown by equation 10. Use of V_p would, however, create a problem because the proportionality factor varies with core dimensions. Thus, to keep no-load flux density within bounds, the bounds on applied voltage would have to vary.

Voltage equation 8 and the temperature rise equations 34 and 35 constitute for $k=1, 2, 3$ a set of nine equations which must be satisfied. These can be written in the form of equation 2:

$$f_{3k-2}=V_p^2-[E_k+R_{pk}(I_s+I_{hk})]^2-[R_{pk}I_{ek}]^2=0 \quad k=1, 2, 3 \quad (44)$$

$$f_{3k-1}=\theta_{ck}-[\theta_{sk}+C_cP_{ck}+C_hP_{hk}]=0 \quad k=1, 2, 3 \quad (45)$$

$$f_{3k}=P_{ck}+P_{hk}-H_{tk}\theta_{sk}=0 \quad k=1, 2, 3 \quad (46)$$

Another equation arises as a result of the requirement that windings operate at the same current density at the design center ambient. Equating J_{p1} and J_{s1} from equation 23 and rearranging terms gives

$$f_{10}=\sigma-I_s/(I_s+I_{p1})=0 \quad (47)$$

Specifications for the ultrahigh-temperature transformer require that

$$R_n\leq R_{ns},\ R_f\leq R_{fs},\ P_{o1}=P_{os},\ \theta_{cs}\leq t_{cs}-t_{ak} \quad (48)$$

where R_{ns} is the maximum allowable no-load regulation; R_{fs} the maximum allowable full-load regulation; P_{os} the specified power output at design center ambient; and t_{cs} the maximum allowable coil temperature. The regulation requirements will be expressed as equations by introducing slack variable z_n and z_f.

$$f_{11}=R_n+Z_n-R_{ns}=0 \quad (49)$$

$$f_{12}=R_f+Z_f-R_{fs}=0 \quad (50)$$

Using equation 13, the power output requirement can be expressed as the constraint equation

$$f_{13}=P_{os}-(E_1-R_{s1}I_s)I_s=0 \quad (51)$$

The temperature requirement will be treated as an upper bound, since θ_{c3} is one of the basic variables, and also as bounds for the other coil temperatures. Thus

$$0\leq\theta_{ck}\leq t_{cs}-t_{ak} \quad k=1, 2, 3 \quad (52)$$

The same bounds, arbitrarily imposed on the surface rise, give

$$0\leq\theta_{sk}\leq t_{cs}-t_{ak} \quad k=1, 2, 3 \quad (53)$$

The range of available excitation data is such that flux density and center-leg width must be restricted in this way:

$$30\leq B_k\leq 100 \quad k=1, 2, 3, 4 \quad (54)$$

$$0.37\leq a\leq 1.50 \quad (55)$$

If the range of data were not limiting, flux density could be allowed to go somewhat higher since saturation density for the core steel is approximately 120 kilolines per square inch.

The secondary conductor area fraction σ must be between zero and one. At the ends of that range, one of the winding resistances will be infinite. Hence, σ will be arbitrarily bounded. Thus

$$0.01\leq\sigma\leq 0.99 \quad (56)$$

To obtain satisfactory mechanical construction, core dimensional ratios will be restricted to the ranges

$$0.5\leq p\leq 3 \quad (57)$$

$$1.0\leq q\leq 5 \quad (58)$$

$$0.5\leq r\leq 3 \quad (59)$$

Load current must be greater than zero, and an upper bound is obtained from the relations

$$I_s=P_{os}/(E_1-R_{s1}I_s)$$

$$R_{s1}I_s<R_{ns}E_1$$

$$E_1>C_EF_ha_l^2r_lB_l$$

wherein subscript l denotes lower bound. Hence

$$0\leq I_s\leq P_{os}/[C_EF_ha_l^2r_lB_l(1-R_{ns})] \quad (60)$$

To meet the specified limits on regulation, the slack variables z_n and z_f must be equal to or greater than zero. Since negative values of R_n and R_f are impossible by definition, the slack variables are restricted. Thus

$$0\leq Z_n\leq R_{ns} \quad (61)$$

$$0\leq Z_f\leq R_{fs} \quad (62)$$

The transformer optimization problem has now been reduced to the form of relations 1 and 2. The variables may be equated to the x_i in any convenient order. There are 18 ($=n$) variables, including the two slack variables, and 13 ($=m$) constraint equations. The criterion function is the transformer weight

$$g=G_T \quad (63)$$

An analytical solution of the problem not being feasible, numerical procedures must be used. However, the general nature of the solution can be deduced. First, study of equation 32 and related equations indicates that the transformer weight increases monotonically with each core dimension and is, therefore, a minimum for a, p, q, and r equal to their lower bounds, the constraint equations being neglected. Second, equation 13 shows that the power output is a quadratic function of load current, and has a maximum with respect to load current and probably also with respect to the other variables. Hence, the optimum design probably is obtained when core dimensions are reduced to the point where either temperature rise, no-load regulation, or full-load regulation is equal to the specified limit, or where the maximum power output for a given set of core dimensions is just equal to the specified output.

Optimization Technique

In the discussion that follows, variables x_i may, for convenience, be regarded as co-ordinates of a point $X=(x_1, \ldots, x_n)$ in n-dimensional space. For brevity, the values of functions f_j and g may be written $f_j(X)$ and $g(X)$. Points that satisfy relations 1 lie in a closed rectangular region R of n-dimensional space; and the locus of points for which $f_j(X)=0$ is a surface S.

The functions f_j and g are assumed to be continuous, finite, and single-valued. The particular functions derived for the transformer optimization possess the assumed properties. A minimum of g can be proved to exist, provided there are points in R for which $f_j(X)=0$. The minimum may be at an interior point where $\alpha_i<x_i<\beta_i$, or at a boundary point where one or more of the $x_i=\alpha_i$ or $=\beta_i$. There may also exist points where g has a local (relative) minimum. (The function g is said to have a local minimum at point X_m if $g(X)>g(X_m)$ at all points in the neighborhood of X_m, the points being restricted to those in region R and surface S.)

Perhaps the most detailed gradient procedure is that developed by Brown.[3] The major difficulty in applying his procedure to the transformer problem is finding variables for which the Jacobian of the functions f_j does not vanish. As mentioned previously, the power output is probably a maximum, at or near the optimum, so it appears the Jacobian will be zero or nearly zero, regardless of which variables are chosen.

Several other methods of solving the general optimization problem are available, perhaps best known of which is by Lagrange multipliers.[4] As this method requires the solution of $(m+n)$ equations, it is rather impractical, except when there are few variables and the functions simple in form. Also, to find minimal points on the boundary requires special treatment. Gradient methods formulate the problem as a system of differential equations to be solved, and convergence properties of such systems have been studied by Arrow, Harwicz, and Uzawa.[5] To obtain solutions of the equations, numerical integration procedures must be used, which for the transformer problem requires a large amount of computation; whether prohibitively large has not been investigated thoroughly.

Since none of the existing methods studied are entirely suitable for the present problem, a quasi-gradient method has been developed and is now presented. Simple in concept, it dictates merely that a sequence of designs be made, each satisfying the constraints, for which the criterion function takes on successively smaller, specified values. The basic technique is to find a point X in region R at which $f_j(X)=0$ and $g(X)\leq\Omega$, where Ω is a parameter having a specified value, by minimizing a certain function ϕ of f_j, g, and Ω.

Assuming for the moment that a procedure exists for finding solutions of the relations

$$f_j(X)=0 \qquad j=1, \ldots, m \tag{64}$$

$$g(X)\leq\Omega \tag{65}$$

then this procedure can be used to find the optimal point X_c, letting X_0 be any point in R for which $f_j(X)=0$. Of course, if no such point exists, the optimization problem has no solution. If the necessary point is found, the next move is to arrive at a sequence of points $[X_r]=[X_1, X_2, \ldots]$ in R that satisfies 64 and 65 with $\Omega=C_r$ and the constants C_r chosen so that

$$C_r<g_c(X_{r-1}) \tag{66}$$

For some values of r, no solution can be found since none exists; hence

$$L_o<g(X_c)\leq U_o \tag{67}$$

where $L_o=C_r$ and $U_o=g(X_{r-1})$. By process of bisection, bounds which are close together, L_1 and U_1, can be found, letting

$$\Omega=(L_o+U_o)/2 \tag{68}$$

If a solution of equation 64 and 65 is found, set $L_1=L_o$ and $U_1=\Omega$; otherwise, set $L_1=\Omega$ and $U_1=U_0$. In either case, $U_1-L_1<U_o-L_o$. Continuing this process, the optimum can be located within an arbitrarily small interval.

Another procedure for solving 64 and 65 is given, but it, unfortunately, does not indicate, if no solution is found, whether one exists. To the author's knowledge, there is no way to determine this. Nevertheless, the process which fails to provide this information must not be considered as worthless, since it may be used to find local minima. Reasonably, if any solutions X_o of 64 can be found and if the difference between $g(X_o)$ and Ω is small, then solutions 64 and 65 can be found, if any exist, because of the continuity of the functions. A function ϕ is to be defined in value as equal to zero if the relations are satisfied, and greater than zero otherwise. Hence, a solution corresponds to an absolute minimum of ϕ.

Before defining ϕ, relation 65 will be changed to an equation by introducing a slack variable x_{n+1} in the manner indicated by equation 39

$$f_{m+1}=g+x_{n+1}-\Omega=0 \tag{69}$$

$$0\leq x_{n+1}\leq\beta_{n+1} \tag{70}$$

where β_{n+1} is any constant $\leq\Omega$.

A suitable definition of ϕ is

$$\phi=\sum_{j=1}^{m+1}(w_jf_j)^2 \tag{71}$$

where the scaling factors w_j are nonzero constants, used to reduce the functions to common units of measure, or to "nondimensional" form. When the value of Ω is fixed, ϕ may be regarded as a function of the $n+1$ variables x_i and its value written $\phi(X)$, where X now denotes a point in $(n+1)$-dimensional space. Region R now includes the restriction 70.

Except for bounds on the variables, ϕ is an unconstrained function and its minimum can be found by gradient methods. Four steps to take in a simple iterative minimization procedure are: (1) select a starting point, (2) find a line through the starting point along which ϕ decreases, (3) find a point on that line where ϕ is approximately a minimum, and (4) repeat, using the approximate minimal point as the starting point of the next iteration.

Making X_o^s the starting point for the $(s+1)$th iteration, let $X_o^s+\lambda\ Y^s$ denote the point having co-ordinates $x_{io}^s+\lambda\ y_i^s$ $(i=1,\ \ldots,\ n+1)$, where λ is a scalar quantity. In the next discussion, superscript s will be omitted except where needed for clarity. If Y is held constant, the point

$$X_\lambda=X_o+\lambda Y \tag{72}$$

traverses a straight line, which passes through X_o as λ varies, and ϕ is a function of the single parameter λ.

To find Y, test the value of ϕ at points near X_o along the co-ordinate axes. Let $\Delta\phi_i$ denote the change in ϕ when x_{io} is changed by an amount Δx_i, and all other variables are held constant. Select a small increment δ_i for each variable, so that

$$|\delta_i|<\beta_i-\alpha_i \tag{73}$$

Now, for each variable in turn, compute $\Delta\phi_i$ for $\Delta x_i=\delta_i$. If $x_{io}+\Delta x_i$ is out of bounds or if $\Delta\phi_i\geq 0$, compute $\Delta\phi_i$ for $\Delta x_i=-\delta_i$. If $x_{io}+\Delta x_i$ is again out of bounds or $\Delta\phi_i\geq 0$, set Δx_i and $\Delta\phi_i$ equal to zero. Next compute

$$D=(\Delta\phi_1)^2+\ldots+(\Delta\phi_{n+1})^2 \tag{74}$$

If $D\neq 0$, then

$$y_i=(\Delta x_i)(\Delta\phi_i)/D \tag{75}$$

If $D=0$, the iteration is terminated: the point X_o^s is then the limit point of the iterative process. Note that y_i is equal to or less than the magnitude of the selected increment.

Having found Y, an approximate minimum of ϕ is determined by successively setting $\lambda=1, 2, \ldots$ until, for some value of λ, either (a) point X_λ is out of bounds, or (b) ϕ increases:

$$\phi(X_\lambda)\geq\phi(X_{\lambda-1}) \tag{76}$$

If (a) occurs first, the approximate minimum is on the boundary, and the boundary point is found by linear interpolation. If (b) occurs first, the approximate minimum is $\phi(X_{\lambda-1})$, and point $X_{\lambda-1}$ is taken as the starting point for the next iteration. In some cases, an improved estimate of the minimal point can be obtained by quadratic interpolation.[3,4] Let $\phi_\lambda=\phi(X_\lambda)$. If $\lambda\geq 2$, compute

$$\mu=\lambda-2+\frac{\phi_\lambda-4\phi_{\lambda-1}+3\phi_{\lambda-2}}{2\phi_\lambda-4\phi_{\lambda-1}+2\phi_{\lambda-2}} \tag{77}$$

Now, if $\phi_\mu<\phi_{\lambda-1}$, take X_μ as the starting point for the next iteration, rather than $X_{\lambda-1}$.

Once the limit point is reached, it may be advisable to repeat the entire process with smaller increments. Under ideal conditions, the limit is in the neighborhood of a point where $\phi=0$. However, it must be recognized that this point may be near a local minimum where $\phi>0$ or a saddle point where ϕ is a maximum with respect to some of the variables. Techniques[3] are available, which will determine whether the limit point is a saddle point, and also outline further steps in the minimization process. Also, it may happen that, after Y is determined, the value of ϕ increases on the first step $(\lambda=1)$ so the computation will "oscillate." This situation is easily detectable, and the computational program can be arranged to set all components of Y, except one, equal to zero.

In solving equations 64 and 65, it is natural to take as the starting point X_o^s, the point X_r augmented by the co-ordinate $x_{n+1}=0$.

A geometrical interpretation of the quasi-gradient solution is shown in Fig. 3

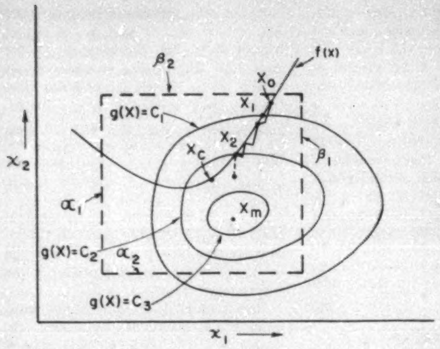


Fig. 3. Optimization process for simple case

for a fictitious case in which there is only one constraint equation and only two variables. In Fig. 3, the line marked $f(X)$ is the locus of points for which $f(X)=0$. The unconstrained minimum of the criterion function is at X_m and the constrained minimum at X_c. The loci of points for which $g(X)=\text{constant}$ are shown as closed contours. The solution process is depicted as a zigzag "trajectory," the significant feature of which tends toward the unconstrained minimum but is "forced" back to the constraint line. (In Brown's procedure, the trajectory would be tangent to the constraint line at X_1, X_2, ...)

Elimination of Variables

Much effort may be expended on minimization procedures which satisfy certain constraint equations, although other available methods might provide an easier solution.

Consider, for example, equation 47, which is simply solved for σ:

$$\sigma=I_s/(I_s+I_{p1}) \tag{78}$$

Hence, σ may be considered a dependent variable and eliminated from the set of basic variables. Whenever the values of function f_f are to be computed, the value of σ must first be computed by 78.

Another example shows how to eliminate the slack variable Z_n. Rearranging equation 49

$$Z_n=R_{ns}-R_n \tag{79}$$

Now, if R_n equals or is less than R_{ns}, equation 79 gives a value for Z_n within the bounds, 61. But, if R_n is greater than R_{ns}, equation 79 gives an inadmissible negative value; in this case, $Z_n=0$ minimizes the function f_{11}^2. Thus, the value of Z_n is obtained quickly. Alternatively, Z_n can be eliminated and function f_{11} redefined as

$$\begin{aligned} f_{11}&=0, \text{ if } R_n\leq R_{ns} \\ &=R_n-R_{ns}, \text{ otherwise} \end{aligned} \tag{80}$$

Other slack variables (Z_f and x_{n+1}) can be eliminated in the same way.

For certain other variables, iteration as explained by Scarborough[7] may be used to obtain numerical values. Thus, equations 44 through 46 may be solved for $B_k, \theta_{ck}, \theta_{sk}$ ($k=1, 2, 3$). This gives

$$B_k=\frac{[V_p^2-(R_{pk}I_{ek})^2]^{1/2}-R_{pk}(I_s+I_{hk})}{C_EA_h} \tag{81}$$

and the original temperature-rise equations 34 and 35. The right-hand sides of these equations are functions of B_k, θ_{ck}, and θ_{sk}. To obtain the solutions, values of B_1, θ_{c1}, and θ_{s1} approximating the solutions are selected, substituted into the equations, and new approximations obtained; this is repeated until the successive approximations differ by less than specified amounts. Solutions for $k=2$ and 3 are obtained similarly, starting from arbitrary values or from solutions obtained previously.

If σ is to be considered a dependent variable along with B_k, θ_{ck}, and θ_{sk}, equation 78 must be included in the iterative process (for $k=1$ only) because I_{p1} is a function of B_1. The process applied to equations 34, 35, 78, and 81 will converge to a solution, except under conditions unlikely to occur, starting from values $B_1=B_4$, $\theta_{c1}=\theta_{s1}=0$, and $\sigma=1/2$.

With variables eliminated by iteration, the next question is how to deal with the bounds on the variables. Defined by relations similar to equation 41, additional constraint functions can be introduced to replace those which were eliminated. Theoretically, a pair of such functions should be defined for each variable, but in practice that is not necessary as some of the bounds will not affect the optimum design. For example, the full-load flux density is less than that of no-load, as may be deduced by inspection of equations 10 and 81; hence, a function corresponding to the upper bound on full-load density is not needed. For temperature rises, the only function needed is

$$\begin{aligned} f&=0, \text{ if } \theta_{cs}\leq t_{cs}-t_{as} \\ &=t_{cs}-t_{as}, \text{ otherwise} \end{aligned} \tag{82}$$

When optimizing the design, if a neglected bound is violated, another constraint function can be introduced and the computational program modified without much difficulty.

Computational Program

A preliminary step is to prepare a generalized subroutine, based on the minimization procedure, for an IBM (International Business Machines Corporation) *650* computer.

For transformer design optimization, the variables are reduced to a, p, q, r, and B_4, (no-load flux density). An iterative solution for the load current

$$I_s=P_{os}/(E_1-R_{s1}I_s) \tag{83}$$

is included. A computer program, based on the quasi-gradient process and incorporating the generalized subroutine, is prepared and checked out.

Computer runs are now made for a 35-volt-ampere rating, starting with the values: $a=0.5$, $p=0.75$, $q=0.25$, $r=1.6875$, and $B_4=100$. Fixed parameters are taken the same as for an existing design; in particular, the conductor fraction F_c, which was 0.13. Transformer weight for those values was 1.05 lb (pounds).

The weight requirement, equation 65, was ignored for the first run, and an acceptable design was produced without changing the values of the basic variables. For the second run, an 0.8-lb requirement was imposed, and again an acceptable design was obtained.

When the requirement was reduced to 0.6 lb, the iterative solution failed to converge and the computer stopped on OVERFLOW. Examination of intermediate results revealed that trying to solve equation 83 by iteration was the source of trouble. This could not be foreseen, because the power output was not known, during the preparation stage of the computer program, to be at or near a maximum for the optimum design. At this writing, the program has not been revised to complete the optimization. Intermediate results indicate that the weight of the optimum design is about 0.5 lb; that the allowable temperature rise is the limiting factor, not regulation; and that the window shape should be more nearly square.

Conclusions

The quasi-gradient optimization technique, when compared with the more sophisticated gradient processes, scores with four points of superiority:

1. It does not require that constraint and criterion functions be differentiable—merely continuous.
2. It is relatively easy to program for a digital computer, and the program can be arranged to produce acceptable designs simply by ignoring the requirement that the criterion function be less than a specified value.
3. It will obtain an improved design without completing the optimization, which is important economically, as well as production-wise.
4. It can be extended to cases where variables are restricted to certain discrete values. (In the minimization procedure previously described, all the variables were changed simultaneously; if some variables are discrete, the function ϕ should be minimized with respect to one discrete variable at a time.)

The computer runs made for a 35-volt-ampere rating demonstrated that the techniques used will produce an optimum ultrahigh-temperature transformer design, the optimum having a weight about 50% less than an existing design—ample proof that computer optimization techniques merit expanded usage.

Nomenclature

V = terminal volts, rms
I = current, rms
J = current density, amp/in.^{2} (amperes per square inch)
I_h = current due to core loss, amp
I_e = magnetizing current, amp
R = winding resistance, ohms
E = induced volts
P_o = output power, watts
P_c = power loss in coil, watts
P_h = core loss, watts
W = core loss, per lb
H = core magnetizing, amp-turns/in.
B = peak flux density, kilolines/sq (square) in.
f = frequency, cps (cycles per second)
ρ = resistivity, ohm-in.
σ = fraction of total conductor area in secondary winding
l_h = mean length magnetic circuit, in.
l_c = mean length of turn, in.
A_h = net cross section of core-center leg, in.^{2}
A_c = total conductor cross-sectional area, in.^{2}
F_h = stacking factor of core
F_c = coil space factor
G_h = weight of core, lb
G_c = weight of conductor, lb
G_e = weight of enclosure, etc., lb
G_T = total transformer weight, lb
δ_h = density of core steel, lb/in.^{3} (cubic in.)
δ_c = density of conductor, lb/in.^{3}
δ_s = density of case steel, lb/in.^{3}
δ_f = density of filling compound, lb/in.^{3}
τ = thickness of case steel, in.
a = core-center leg width, in.
b = core-window width, in.
c = core-window height, in.
d = core-stack height, in.
p = ratio, window width to center-leg width
q = ratio, window length to width
r = ratio, stack height to center-leg width
t_a = ambient temperature, C
T_a = ambient temperature, K (degrees Kelvin)
θ_c = temperature rise of coil, C
θ_s = temperature rise of transformer surface, C
F_A, C_c, C_h, C_1, C_2 = parameters in temperature-rise equations
H_t = surface heat transfer coefficient, watts/C
α = thermal coefficient of resistivity, 1/C
R_n = regulation, no load to full load
R_f = regulation, full load over ambient temperature range
R_{ns} = maximum allowable no-load regulation
R_{fs} = maximum allowable full-load regulation
t_{cs} = maximum allowable coil temperature, C
P_{os} = specified volt-ampere output

Subscripts $_p$, $_s$, and $_k$ refer, respectively, to primary winding, secondary winding, and operating condition.

References

1. ULTRAHIGH TEMPERATURE ELECTRONIC TRANSFORMERS; PART I. DESIGN CONSIDERATIONS, **J. F. Rippin, Jr., H. B. Harms, Glenn Walters.** *AIEE Transactions*, see pp. 295–302 of this issue.
2. ULTRAHIGH TEMPERATURE POWER TRANSFORMERS AND INDUCTORS, **H. B. Harms, J. C. Fraser.** *WADC Technical Report 57-492*, Wright Air Development Center, Dayton, Ohio, vol. I, May 1958, pp. 48–51, figs. 20 a, b, c, d.
3. A GENERALIZED COMPUTER PROCEDURE FOR THE DESIGN OF OPTIMUM SYSTEMS—PARTS I AND II, **R. Roderick Brown.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, July 1959, pp. 285–93.
4. HIGHER MATHEMATICS FOR ENGINEERS AND PHYSICISTS (book), **Ivan S. Sokolnikoff, Elizabeth S. Sokolnikoff.** McGraw-Hill Book Company, Inc., New York, N. Y., 1941, pp. 163–65.
5. STUDIES IN LINEAR AND NONLINEAR PROGRAMMING (book), **K. J. Arrow, L. Harwicz, H. Uzawa.** Stanford University Press, Stanford, Calif., 1958, pp. 165–76.
6. NUMERICAL METHODS (book), **A. D. Booth.** Butterworth's Scientific Publications, London, England, 1955, pp. 154–60.
7. NUMERICAL MATHEMATICAL ANALYSIS (book), **J. B. Scarborough.** The Johns Hopkins Press, Baltimore, Md., third edition, 1955, pp. 206–10.

Linear Analysis of the Electromechanical Pulser

H. C. PANDE
ASSOCIATE MEMBER AIEE

WITH THE increasingly higher peak power requirements for radar, thermonuclear research, and other purposes, the electromechanical generation of such pulsed power is becoming more and more important. An electromechanical pulse generator or simply an electromechanical pulser, as such a device may be called, can take a variety of possible forms.[1] However, generally any of the possible forms can be reduced electrically to a system of two or more coils, connected both inductively and conductively. Inductive coupling is basic to the functioning of the device, whereas conductive coupling can help to increase the useful power output.

In its primitive form the electromechanical pulser consists of two coils that are inductively coupled, only one of which is excited; see Fig. 1. The coils are capable of motion relative to one another, causing the mutual inductance between them to vary.

The excitation may also be by means of permanent magnets. In a rotating device the excited or primary coil would be on the stator and the secondary coil on the rotor. As the rotor revolves, the mutual inductance between the two coils would vary periodically with rotor position.

Linear lumped parameter analysis of the electromechanical pulser shows that it is possible under ideal assumptions to feed resistive loads with "modified-impulse" currents. A modified-impulse function is defined as a function, which like an im-

---

Paper 61-100, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 31, 1960; made available for printing December 30, 1960.

H. C. PANDE is with the Microwave Research Institute, Brooklyn, N. Y.

The author thankfully acknowledges the valuable advice and suggestions of Dr. E. Levi of the Polytechnic Institute of Brooklyn. This paper was made possible through the support and sponsorship extended by the Rome Air Development Center under Contract no. *AF-30(602)-2149*.

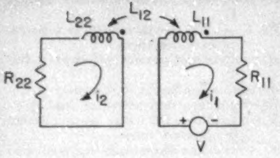


Fig. 1 (left). Circuit diagram

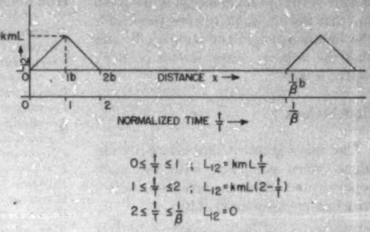


Fig. 2 (right). Mutual inductance versus distance

pulse function, is infinite at one point and has a finite integral. However, unlike an impulse function it may have finite nonzero values elsewhere.

This highly significant effect is due to the critical choice of parameters which, in the ideal case, cause the flux density at the leading edge of the moving coil to build up without bound.

Piecewise Linear Analysis

The mutual inductance between the coils in the primitive arrangement is a function of position; for the purpose of analysis let it be considered piecewise linear with respect to the space coordinate x, as in Fig. 2. The mutual inductance is periodic in x with the period $1/\beta$. Let the subscripts 1 and 2 refer to the primary, viz., the excited coil and the secondary respectively.

The electrical network equations and the mechanical equation of motion governing the system are the following:

$$\frac{d}{dt}(L_{11}i_1+L_{12}i_2)+R_{11}i_1=V \qquad (1A)$$

$$\frac{d}{dt}(L_{12}i_1+L_{22}i_2)+R_{22}i_2=0 \qquad (1B)$$

$$m\frac{d^2x}{dt^2}-\frac{\partial L_{12}}{\partial x}(i_1i_2)+f=F \qquad (1C)$$

Due to the presence of a product term in the variables on the left-hand side of equation 1(C), the system is essentially nonlinear. However, if the mechanical motion is specified as a time function, then the equations requiring solutions are 1(A) and 1(B) which become linear with time-dependent coefficients. The relative motion between the coils will be assumed constant and limiting cases will be considered. In all cases $L_{11}=L_{22}=L$.

Since constant relative velocity has been assumed, the space axis can be converted into the time axis by means of a linear transformation. It is useful to normalize the time axis in terms of the time of approach T defined as the time required to traverse the distance b or the time to reach the maximum coupling from the initial zero coupling.

Case 1. No Dissipation in Either Coil

$R_{11}=R_{22}=0$

Equations 1(A) and 1(B) reduce to

$$\frac{d}{dt}(Li_1+L_{12}i_2)=0 \qquad (2A)$$

$$\frac{d}{dt}(L_{12}i_1+Li_2)=0 \qquad (2B)$$

Initial conditions at $t=0$

primary flux linkages $=\lambda_{10}$
secondary flux linkages $=0$

The corresponding solutions are

$$i_1=\frac{I_{10}}{1-k^2} \qquad (3A)$$

$$i_2=-\frac{I_{10}k}{1-k^2} \qquad (3B)$$

Since the coefficient of coupling k is a function of position only, the currents are also only a function of position, quite independent of the velocity. Clearly, if unity coupling or zero leakage were possible, the coil currents would tend to increase without bound as the coupling approaches unity. Since both the coils have zero resistance the flux lines can neither leave the primary nor enter the secondary coil. Thus the flux is forced between the coils, causing high flux density to develop at the leading edge of the moving coil. This induces a sheet current density equal in magnitude to the magnetic-field intensity tangential to the boundaries of the coils. In the ideal case of unity coupling the entire initial flux is forced into an area that has decreased to zero. This causes an unbounded flux density and results in unbounded currents.

Case 2. Primary and Secondary Coils Equally Dissipative

$R_{11}=R_{22}=R$, Let $\tau=\dfrac{L}{R}$ and $I_{10}=\dfrac{V}{R}$

The solution in this case has to be in three parts. For the region $0\leq t/T\leq 1$, equations 1(A) and 1(B) reduce to

$$\tau\frac{di_1}{dt}+k_m\tau\frac{t}{T}\frac{di_2}{dt}+k_m\frac{\tau}{T}\,i_2+i_1=I_{10} \qquad (4A)$$

$$\tau\frac{di_2}{dt}+k_m\tau\frac{t}{T}\frac{di_1}{dt}+k_m\frac{\tau}{T}i_1+i_2=0 \qquad (4B)$$

Initial conditions at $t=0$

$i_1=I_{10}$ and $i_2=0$

The corresponding solutions are for $T/k_m\tau\neq 1$

$$i_1=I_{10}\left[\frac{\left(\frac{T}{k_m\tau}\right)^2}{\left(\frac{T}{k_m\tau}\right)^2-1}+\frac{\left(1-k_m\frac{t}{T}\right)^{\frac{T}{k_m\tau}-1}}{2\left(1-\frac{T}{k_m\tau}\right)}+\frac{\left(1+k_m\frac{t}{T}\right)^{-\frac{T}{k_m\tau}-1}}{2\left(1+\frac{T}{k_m\tau}\right)}\right] \qquad (5A)$$

$$i_2=I_{10}\left[\frac{\frac{T}{k_m\tau}}{1-\left(\frac{T}{k_m\tau}\right)^2}-\frac{\left(1-k_m\frac{t}{T}\right)^{\frac{T}{k_m\tau}-1}}{2\left(1-\frac{T}{k_m\tau}\right)}+\frac{\left(1+k_m\frac{t}{T}\right)^{-\frac{T}{k_m\tau}-1}}{2\left(1+\frac{T}{k_m\tau}\right)}\right] \qquad (5B)$$

and for $T/k_m\tau=1$

$$i_1=I_{10}\left[0.75-0.5\ln\left(1-k_m\frac{t}{T}\right)+0.25\left(1+k_m\frac{t}{T}\right)^{-2}\right] \qquad (6A)$$

$$i_2=I_{10}\left[-0.25+0.5\ln\left(1-k_m\frac{t}{T}\right)+0.25\left(1+k_m\frac{t}{T}\right)^{-2}\right] \qquad (6B)$$

Critical Values of the Dissipation Factor

One may define T/τ as the dissipation factor. If the dissipation factor is such that $T/k_m\tau\leq 1$, then an examination of the second terms in the expressions for the currents in the coils in equations 5 and 6 leads to a result which, from the point of view of pulse generators, is highly significant and quite unexpected. In the ideal case when the maximum coefficient of coupling k_m is unity, it is at once apparent that with the dissipation factor within the upper bound, as mentioned above, the currents increase without bound as $t/T \to 1$. This is a remarkable result considering the fact that both coils are dissipative. Furthermore, it must be noted that for a dissipation factor such that $T/k_m\tau > 1$, the currents remain bounded even with unity coupling.

Clearly, the dissipation factor has a critical value which determines whether in the ideal case of unity coupling the coil currents would be bounded or not. Quite obviously in the ideal case the critical dissipation factor is unity.

As both coils are dissipative the initial stored magnetic flux can diffuse through both the primary and the secondary coil. However, when the dissipation factor has the value $T/k_m\tau \leq 1$, the decrease in area into which the remaining flux (initial flux less the flux that has diffused through the two coils) is forced is such that the resulting flux density increases. In the ideal case there is finite nonzero flux that is forced into zero area causing unbounded flux density and unbounded currents. If the dissipation factor is greater than unity, at unity coupling the remaining flux is zero. Obviously, flux density is not unbounded in this case and no unbounded currents can arise.

On first thought, the existence of a critical dissipation factor that causes the currents to be unbounded does not appear extraordinarily interesting, since if the energy required for generating such currents is infinite, then clearly such currents cannot be generated even in the ideal case.

To check for the boundedness of the energy E, associated with the current pulse, it is only necessary to calculate it for the current components that increase without bound, viz., the second terms on the right-hand side of equations 5. For these, one obtains for the ideal case ($k_m = 1$)

$$E = \int_0^T i^2 R\,dt = \int_0^T \frac{R\left(1-\frac{t}{T}\right)^{-2\left(1-\frac{T}{\tau}\right)}}{4\left(1-\frac{T}{\tau}\right)^2}\,dt$$

$$= \frac{RT}{4\left(1-\frac{T}{\tau}\right)^2}\left[\frac{\left(1-\frac{t}{T}\right)^{-1+2\frac{T}{\tau}}}{2\frac{T}{\tau}-1}\right]_T^0 \qquad (7)$$

Another critical value of the dissipation factor is now evident. For $T/\tau < 1/2$

$\int_0^T i^2 R\,dt = \text{unbounded}$

Thus $1/2 < T/\tau \leq 1$ is the most interesting range: with the dissipation factor within this range the currents become unbounded at $t/T = 1$, the energy required is finite, and the time integral of the currents is also finite. It follows that this corresponds to the modified-impulse function.

In the ideal case ($k_m = 1$), the importance of the critical values of the dissipation factor is unmistakably clear. It has to be between the two values in order to generate modified-impulse currents with finite energies.

The critical values of the dissipation factor are of theoretical interest only, for unity coefficient of coupling, or zero leakage, is not physically attainable; hence the ideal case cannot be realized in practice. Consequently, in the practical case, the coil currents as well as the associated energy remain bounded for any dissipation factor. Seen from the practical standpoint the dissipation factor still retains its importance; it is evident from equations 5 that the lower the dissipation factor T/τ, the higher the peak magnitude of the currents in the coils.

CASE 2 CONTINUED

For the region $1 \leq t/T \leq 2$, equations 1(A) and 1(B) reduce to

$$\tau\frac{di_1}{dt} + k_m\tau\left(2-\frac{t}{T}\right)\frac{di_2}{dt} - k_m\frac{\tau}{T}i_2 + i_1 = I_{10} \qquad (8A)$$

$$\tau\frac{di_2}{dt} + k_m\tau\left(2-\frac{t}{T}\right)\frac{di_1}{dt} - k_m\frac{\tau}{T}i_1 + i_2 = 0 \qquad (8B)$$

The corresponding solutions are, for $T/k_m\tau \neq 1$

$$i_1 = \frac{I_{10}\left(\frac{T}{k_m\tau}\right)^2}{\left(\frac{T}{k_m\tau}\right)^2-1} + \frac{c_1}{2}\left(1+2k_m-k_m\frac{t}{T}\right)^{\frac{T}{k_m\tau}-1} + \frac{c_2}{2}\left(1-2k_m+k_m\frac{t}{T}\right)^{-1-\frac{T}{k_m\tau}} \qquad (9A)$$

$$i_2 = \frac{I_{10}\frac{T}{k_m\tau}}{\left(\frac{T}{k_m\tau}\right)^2-1} + \frac{c_1}{2}\left(1+2k_m-k_m\frac{t}{T}\right)^{\frac{T}{k_m\tau}-1} - \frac{c_2}{2}\left(1-2k_m+k_m\frac{t}{T}\right)^{-1-\frac{T}{k_m\tau}} \qquad (9B)$$

and for $T/k_m\tau = 1$

$$i_1 = \frac{c_1'}{2} - \frac{I_{10}}{2}\ln\left(1+2k_m-k_m\frac{t}{T}\right) + \frac{I_{10}}{4} + \frac{c_2}{2}\left(1-2k_m+k_m\frac{t}{T}\right)^{-2} \qquad (10A)$$

$$i_2 = \frac{c_1'}{2} - \frac{I_{10}}{2}\ln\left(1+2k_m-k_m\frac{t}{T}\right) - \frac{I_{10}}{4} - \frac{c_2}{2}\left(1-2k_m+k_m\frac{t}{T}\right)^{-2} \qquad (10B)$$

Since the currents i_1 and i_2 are continuous at $t/T = 1$, the solutions in the two regions can be matched at that point to obtain the value of the constants involved.

For the region $2 \leq t/T \leq 1/\beta$, equations 1(A) and 1(B) reduce to

$$\tau\frac{di_1}{dt} + i_1 = I_{10} \qquad (11A)$$

$$\tau\frac{di_2}{dt} + i_2 = 0 \qquad (11B)$$

The solutions are evident:

$$i_1 = c_3 e^{-\frac{t}{\tau}} + I_{10} \qquad (12A)$$

$$i_2 = c_4 e^{-\frac{t}{\tau}} \qquad (12B)$$

Again the continuity of i_1 and i_2 at $t/T = 2$ can be utilized to obtain the needed constants.

Equations 5, 9, and 12, and 6, 10, and 12 give the required solutions for the entire region.

CASE 3. PRIMARY COIL DISSIPATIVE, SECONDARY DISSIPATIONLESS

$R_{11} = R$, $R_{22} = 0$, Let $\tau = \frac{L}{R}$ and $I_{10} = \frac{V}{R}$

For the region $0 \leq t/T \leq 1$ equations 1(A) and 1(B) reduce to

$$\tau\frac{di_1}{dt} + k_m\tau\frac{t}{T}\frac{di_2}{dt} + k_m\frac{\tau}{T}i_2 + i_1 = I_{10} \qquad (13A)$$

$$\frac{d}{dt}\left(Li_2 + k_m L\frac{t}{T}i_1\right) = 0 \qquad (13B)$$

Initial conditions at

$t = 0$

$i_1 = I_{10}$ and $i_2 = 0$

The corresponding solutions are

$$i_1 = I_{10}\frac{\left(1-k_m\frac{t}{T}\right)^{\frac{T}{2k_m\tau}-1}}{\left(1+k_m\frac{t}{T}\right)^{1+\frac{T}{2k_m\tau}}} \times \left[\frac{1}{\tau}\int\left(\frac{1+k_m\frac{t}{T}}{1-k_m\frac{t}{T}}\right)^{\frac{T}{2k_m\tau}}dt + c_5\right] \qquad (14A)$$

$$i_2 = -k_m\frac{t}{T}i_1 \qquad (14B)$$

Unless $T/2k_m\tau$ is an integer, the integral involved in the expression for the primary current cannot be solved in a closed form and a series solution is necessary.

For fractional $T/2k_m\tau$, a series solution follows:

$$i_1 = \frac{I_{10}c_5\left(1-k_m\frac{t}{T}\right)^{\frac{T}{2k_m\tau}-1}}{\left(1+k_m\frac{t}{T}\right)^{1+\frac{T}{2k_m\tau}}} + I_{10}\frac{T}{k_m\tau\left(1+k_m\frac{t}{T}\right)}\times \left[\frac{1}{\frac{T}{2k_m\tau}-1} - \frac{\frac{T}{2k_m\tau}}{\left(\frac{T}{2k_m\tau}-1\right)\left(\frac{T}{2k_m\tau}-2\right)}\times \left(\frac{1-k_m\frac{t}{T}}{1+k_m\frac{t}{T}}\right) + \frac{\frac{T}{2k_m\tau}}{\left(\frac{T}{2k_m\tau}-2\right)\left(\frac{T}{2k_m\tau}-3\right)}\times \left(\frac{1-k_m\frac{t}{T}}{1+k_m\frac{t}{T}}\right)^2 + \ldots\right] \tag{15}$$

$$\frac{T}{k_m\tau} \neq \text{integer}$$

The constant c_5 can be obtained from the initial condition, namely, at $t=0$, $i_1=I_{10}$.

Critical Values of the Dissipation Factor

Considering the ideal case when the maximum coefficient of coupling k_m is unity, it is evident that the first term in the expression for the primary current in equation 15 would increase without bound as t/T approaches unity if $T/\tau<2$. This is the first critical value of the dissipation factor.

The second critical value is obtained, as before, by considering the energy integral. Only the energy component that tends to increase most rapidly need be considered:

$$E=\int_0^T i^2 R\,dt = \left[\frac{RT\left(1-\frac{t}{T}\right)^{\frac{T}{\tau}-1}}{\left(\frac{T}{\tau}-1\right)\left(1+\frac{t}{T}\right)^{\frac{T}{\tau}+2}}\right]_T^0 + \text{slower increasing terms} \tag{16}$$

It is evident that the associated energy would be bounded if $T/\tau>1$.

It follows that the range of the dissipation factor which leads to the generation of modified-impulse currents in the ideal case with finite energies has to be

$$1<\frac{T}{\tau}<2$$

The critical dissipation factor values obtained here are twice those of case 2 where both coils were dissipative. It may be stated here that it is possible to redefine τ in terms of the system rather than in terms of an individual coil. One may define τ as the ratio of the total initial inductance of the system to the total resistance. With this definition the critical range becomes the same in both cases, viz., $1/2<T/\tau<1$.

As in case 2, the currents in the coils as well as the associated energy would remain bounded for any dissipation factor in the practical case because of the unrealizability of unity coupling. It is evident also that the lower the dissipation factor, the higher the peak magnitude of the currents.

CASE 3 CONTINUED

For the region $1\leq t/T\leq 2$ the equations 1(A) and 1(B) reduce to

$$\tau\frac{di_1}{dt}+k_m\tau\left(2-\frac{t}{T}\right)\frac{di_2}{dt}-k_m\frac{\tau}{T}i_2+i_1=I_{10} \tag{17A}$$

$$\frac{d}{dt}\left[Li_2+k_mL\left(2-\frac{t}{T}\right)i_1\right]=0 \tag{17B}$$

The corresponding solutions are

$$i_1=I_{10}\frac{\left[1+k_m\left(2-\frac{t}{T}\right)\right]^{\frac{T}{2k_m\tau}-1}}{\left[1-k_m\left(2-\frac{t}{T}\right)\right]^{\frac{T}{2k_m\tau}+1}}\times\left\{\frac{1}{\tau}\int\left[\frac{1-k_m\left(2-\frac{t}{T}\right)}{1+k_m\left(2-\frac{t}{T}\right)}\right]^{\frac{T}{2k_m\tau}}dt+c_6\right\} \tag{18A}$$

$$i_2=-k_m\left(2-\frac{t}{T}\right)i_1 \tag{18B}$$

As before, for fractional $T/2k_m\tau$, a series solution follows:

$$i_1=\frac{I_{10}c_6\left[1+k_m\left(2-\frac{t}{T}\right)\right]^{\frac{T}{2k_m\tau}-1}}{\left[1-k_m\left(2-\frac{t}{T}\right)\right]^{\frac{T}{2k_m\tau}+1}}+$$

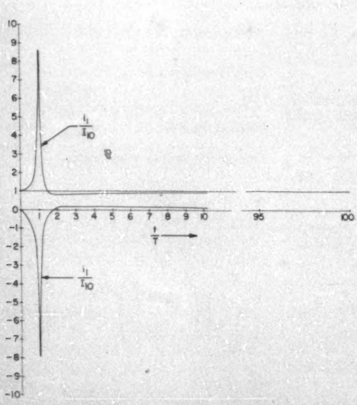


Fig. 3. Pulse shape; equally dissipative case, $k_m=0.95$, $T/k_m\tau=0.1$ and $1/\beta=100$

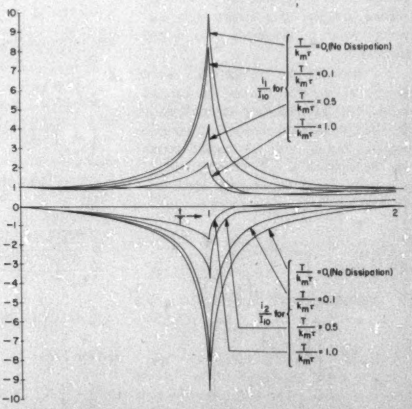


Fig. 4. Enlarged pulse shapes for both coils equally dissipative, $k_m=0.95$

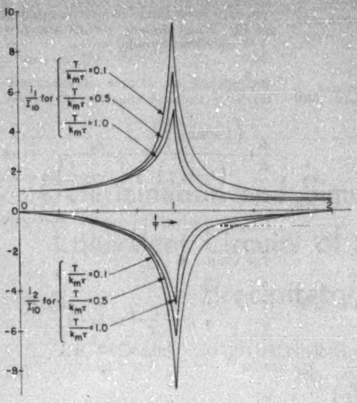


Fig. 5 (left). Enlarged pulse shapes for primary dissipative and secondary lossless, $k_m=0.95$

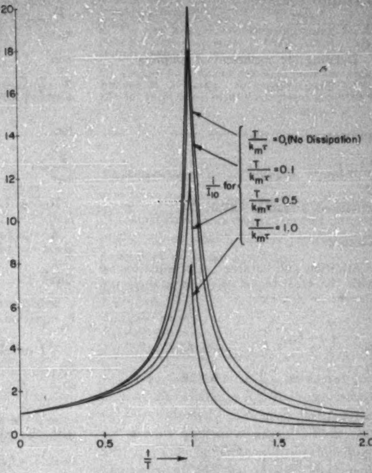


Fig. 6 (right). Enlarged pulse shapes for conductive coupling between the coils, $k_m=0.95$

$$\frac{I_{10}\dfrac{T}{k_m\tau}}{\left[1+k_m\left(2-\dfrac{t}{T}\right)\right]}\times$$

$$\left\{\frac{1}{1+\dfrac{T}{2k_m\tau}}-\frac{\dfrac{T}{2k_m\tau}}{\left(1+\dfrac{T}{2k_m\tau}\right)\left(2+\dfrac{T}{2k_m\tau}\right)}\times\right.$$

$$\left[\frac{1-k_m\left(2-\dfrac{t}{T}\right)}{1+k_m\left(2-\dfrac{t}{T}\right)}\right]+$$

$$\frac{\dfrac{T}{2k_m\tau}}{\left(2+\dfrac{T}{2k_m\tau}\right)\left(3+\dfrac{T}{2k_m\tau}\right)}\times$$

$$\left.\left[\frac{1-k_m\left(2-\dfrac{t}{T}\right)}{1+k_m\left(2+\dfrac{t}{T}\right)}\right]^2+\ldots\right\} \quad (19)$$

As the currents are continuous at $t/T=1$, the solution for the two regions can be matched at that point to obtain the required constant.

For the region $2\leq t/T\leq 1/\beta$, the solutions are exactly as in case 2.

$$i_1=c_7e^{-\frac{t}{\tau}}+I_{10} \quad (20A)$$

$$i_2=c_8e^{-\frac{t}{\tau}} \quad (20B)$$

Continuity requirement at $t/T=2$ gives the needed constants. From equation 18(B) clearly $c_8=0$.

Equations 15, 19, and 20 give the required solution for the entire region.

Conclusions

The theoretical realizability of electromechanical impulse generators capable of feeding finite resistive loads that follows from the foregoing analysis clearly serves to stress the importance of the device as a pulse generator in the practical case of finite nonzero leakage. As far as the design of pulse generators is concerned, the most important single criterion appears to be the dissipation factor, which, as stated earlier, is an index of the peak current. Although lower dissipation factor implies higher peak current it does not follow that the dissipation factor should be kept at the minimum possible value irrespective of other considerations.

From the ultimate design viewpoint, the magnetic leakage and the time of approach must be kept at the minimum possible value, whereas the initial stored magnetic energy must be kept at the maximum possible value. However, the load resistance must be determined from the efficiency consideration and must be many times the unavoidable coil resistance to obtain higher efficiency.

Figs. 3 through 6 show the plot of the primary and the secondary currents in time. As one would expect, the nondissipative case gives the highest pulse. For low dissipation factors, the pulse generated does not differ much from the pulse obtained in the nondissipative case. The generated current pulse is triangular in shape. Since the pulse shape is primarily determined by the mutual inductance variation with distance, it is possible to generate different pulse shapes e.g., a flat-topped pulse, by properly designing the mutual inductance variation.

Nomenclature

c, c', c_1 through c_8 = constants
E = energy associated with the current pulse
F = mechanical driving force
f = frictional force
I_{10} = initial primary current or equivalent = $V/R=\lambda_{10}/L$
i_1 = primary current
i_2 = secondary current
k = coefficient of coupling = $L_{12}/\sqrt{L_{11}L_{22}}$
k_m = maximum coefficient of coupling
L_{11} = primary self inductance = L
L_{12} = mutual inductance
L_{22} = secondary self inductance = L
m = mass of the moving part of the system
R_{11} = primary resistance
R_{22} = secondary resistance
T = time of approach
t = time
V = source voltage in the primary circuit
x = space co-ordinate
β = duty cycle
λ_{10} = initial primary flux linkage
$\tau=L/R$ (inductance reactance)

Appendix I. Solution of Equations

Case 1

On integration equations 2 give

$$Li_1+L_{12}i_2=\lambda_{10}$$

$$L_{12}i_1+Li_2=0 \quad (21)$$

which yields for the currents

$$i_1=\frac{\lambda_{10}L}{L^2-L_{12}^2}$$

$$i_2 = -\frac{\lambda_{10} L_{12}}{L^2 - L_{12}^2} \tag{22}$$

Defining the coefficient of coupling

$$k = \frac{L_{12}}{\sqrt{L_{11} L_{22}}} = \frac{L_{12}}{L}$$

and recognizing that

$$\frac{\lambda_{10}}{L} = I_{10}$$

equations 3 readily follow.

Case 2

Addition and subtraction of equations 4 with the substitution of $i^+ = i_1 + i_2$ and $i^- = i_1 - i_2$ yields

$$\frac{di^+}{dt} + \frac{\left(1 + k_m \frac{\tau}{T}\right)}{\tau\left(1 + k_m \frac{t}{T}\right)} i^+ = \frac{I_{10}}{\tau\left(1 + k_m \frac{t}{T}\right)}$$

$$\frac{di^-}{d} + \frac{\left(1 - k_m \frac{\tau}{T}\right)}{\tau\left(1 - k_m \frac{t}{T}\right)} i^- = \frac{I_{10}}{\tau\left(1 - k_m \frac{t}{T}\right)} \tag{23}$$

with the initial conditions at

$t = 0$

$i^+ = i^- = I_{10}$

The differential equations 23 are linear first order of the standard type

$$\frac{dy}{dx} + Py = Q$$

In terms of the integrating factor $I = e^{\int P dx}$ this has the solution

$$yI = \int QI dx + \text{constant}$$

The appropriate integrating factors are

$$e^{\int \frac{\left(1 + k_m \frac{\tau}{T}\right)}{\tau\left(1 + k_m \frac{t}{T}\right)} dt} = \left(1 + k_m \frac{t}{T}\right)^{1 + \frac{T}{k_m \tau}}$$

$$e^{\int \frac{\left(1 - k_m \frac{\tau}{T}\right)}{\tau\left(1 - k_m \frac{t}{T}\right)} dt} = \left(1 - k_m \frac{t}{T}\right)^{1 - \frac{T}{k_m \tau}} \tag{24}$$

The solutions follow:

$$i^+ = I_{10} \times \left[\frac{\frac{T}{k_m \tau}}{1 + \frac{T}{k_m \tau}} + \frac{1}{\left(1 + \frac{T}{k_m \tau}\right)\left(1 + k_m \frac{t}{T}\right)^{1 + \frac{T}{k_m \tau}}}\right]$$

$$i^- = I_{10} \times \left[\frac{\frac{T}{k_m \tau}}{\frac{T}{k_m \tau} - 1} - \frac{1}{\left(\frac{T}{k_m \tau} - 1\right)\left(1 - k_m \frac{t}{T}\right)^{1 - \frac{T}{k_m \tau}}}\right] \tag{25}$$

for

$$\frac{T}{k_m \tau} \neq 1$$

and

$$i^- = I_{10}\left[1 - \ln\left(1 - k_m \frac{t}{T}\right)\right] \tag{26}$$

for

$$\frac{T}{k_m \tau} = 1$$

Recognizing that $i_1 = (i^+ + i^-)/2$ and $i_2 = (i^+ - i^-)/2$, equations 5 follow.

A similar procedure with equations 8 gives the solutions in equations 9.

Case 3

Substitution for i_2 from equation 13(B) in equation 13(A) gives

$$\frac{di_1}{dt} + \frac{\left(1 - 2\tau \frac{k_m^2}{T} t\right) i_1}{\tau\left(1 - k_m^2 \frac{t^2}{T^2}\right)} = \frac{I_{10}}{\tau\left(1 - k_m^2 \frac{t^2}{T^2}\right)} \tag{27}$$

with the initial condition at

$t = 0, \ i_1 = I_{10}$

The appropriate integrating factor is

$$e^{\int \frac{\left(1 - 2\tau \frac{k_m^2}{T^2} t\right) dt}{\tau\left(1 - k_m^2 \frac{t^2}{T^2}\right)}} = \left(1 - k_m^2 \frac{t^2}{T^2}\right) \times \left(\frac{1 + k_m \frac{t}{T}}{1 - k_m \frac{t}{T}}\right)^{\frac{T}{2k_m \tau}} \tag{28}$$

The solution in equation 14(A) follows. Exactly the same procedure with equations 17 gives the solutions in equations 18.

Appendix II. Conductive Coupling

An addition of conductive coupling in the primitive arrangement would mean that both the coils are suitably connected in series. The electric circuit would then appear as in Fig. 7.

Equations 1(A) and 1(B) now reduce to

$$2L \frac{di}{dt} - 2\frac{d}{dt}(L_{12} i) + Ri = V \tag{29}$$

With the assumption of constant velocity equation 29 becomes first-order linear with time-varying coefficients. Assuming the piecewise linear mutual inductance variation of Fig. 2 and defining $\tau = L/R$ and $I_{10} = V/R$ equation 29 yields, for

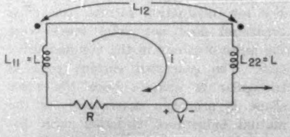


Fig. 7. Circuit diagram with conductive coupling

$$0 \leq \frac{t}{T} \leq 1$$

$$\frac{di}{dt} + \frac{\left(1 - 2k_m \frac{\tau}{T}\right) i}{2\tau\left(1 - k_m \frac{t}{T}\right)} = \frac{I_{10}}{2\tau\left(1 - k_m \frac{t}{T}\right)} \tag{30}$$

$$\text{Integrating factor} = \left(1 - k_m \frac{t}{T}\right)^{1 - \frac{T}{2k_m \tau}} \tag{31}$$

The solution is

$$i = \frac{I_{10} \frac{T}{2k_m \tau}}{\frac{T}{2k_m \tau} - 1} + \frac{I_{10}}{1 - \frac{T}{2k_m \tau}} \left(1 - k_m \frac{t}{T}\right)^{\frac{T}{2k_m \tau} - 1} \tag{32}$$

for

$$\frac{T}{2k_m \tau} \neq 1$$

and

$$i = I_{10} - I_{10} \ln\left(1 - k_m \frac{t}{T}\right) \tag{33}$$

for

$$\frac{T}{2k_m \tau} = 1$$

The critical dissipation factor range is clearly the same as that of case 3, viz., $1 \leq T/\tau \leq 2$. Furthermore, for high maximum coefficient of coupling k_m and low-dissipation factors, a comparison of the last term in equation 32 with the second terms on the right-hand side of equation 5 and the first term on the right-hand side of equation 15 shows that with conductive coupling the peak current is nearly twice that of the other two cases. Quite clearly the peak power capability of the device is increased four times.

For the region $1 \leq t/T \leq 2$ equation 29 reduces to

$$\frac{di}{dt} + \frac{\left(1 + 2k_m \frac{\tau}{T}\right) i}{2\tau\left[1 - 2k_m + k_m \frac{t}{T}\right]} = \frac{I_{10}}{2\tau\left[1 - 2k_m + k_m \frac{t}{T}\right]} \tag{34}$$

$$\text{Integrating factor} = \left(1 - 2k_m + k_m \frac{t}{T}\right)^{1 + \frac{T}{2k_m \tau}}$$

The solution is

$$i = \frac{I_{10} \frac{T}{2k_m \tau}}{1 + \frac{T}{2k_m \tau}} + \frac{c}{\left(1 - 2k_m + k_m \frac{t}{T}\right)^{1 + \frac{T}{2k_m \tau}}} \tag{35}$$

The constant can be obtained by utilizing the continuity of the current at $t/T = 1$.

For the region $2 \leq t/T \leq 1/\beta$ the solution is evident:

$$i = I_{18} + c' e^{-\frac{t}{\beta\tau}} \tag{36}$$

The constant c' can be obtained by considering the continuity condition on current at $t/T = 2$. Equations 32, 33, 35, and 36 give the required solution for the entire range.

References

1. Electromechanical Pulse Generators, E. Levi. *Research Report R-751-59, PIB-679*, Microwave Research Institute, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1959.

2. On the Electromechanical Pulser, H. C. Pande. *Memorandum 16, PIBMRI-833-60*, Polytechnic Institute of Brooklyn, 1960.

3. Linear Analysis of the Electromechanical Pulser, H. C. Pande. *Research Report PIBMRI-876-60*, Polytechnic Institute of Brooklyn, 1960.

Determination and Simulation of the Equivalent Circuits of Electrostatic Precipitators

J. B. THOMAS
MEMBER AIEE

J. W. DRENNING
MEMBER AIEE

H. T. WILLIAMS
NONMEMBER AIEE

CORONA FIELD in the usual electrostatic precipitator is energized by half- or full-wave rectifiers operating at line frequencies, the rectifier output being filtered by the capacitance of the precipitator itself. Analysis of the equivalent circuits involved is somewhat complicated. The rectifier, by nature nonlinear, must be considered as well as the nonlinear electrical characteristic of the corona field. Hence, only certain easily measured quantities, such as average current and voltage, customarily enter into a determination of performance of such circuits.

The foregoing procedure is convenient, but it is not immune from dangers of oversimplification and misinterpretation of precipitator behavior. A more detailed treatment not only provides a realistic understanding of the problems but allows, for example, laboratory simulation of the circuits in question. The latter is often of great practical importance and usefulness when designing and developing equipment for the electrostatic precipitator.

In this paper, equivalent circuits for such precipitators are presented and examined, simulation is discussed, and practical realizations are given. Performance of the realizations is compared with that of field installations. Steady-state analysis only is considered, although problems of transient behavior are equally important, since they affect such things as transformer and regulator design, and operating stability.

The Equivalent Circuit

A typical electrostatic precipitator is illustrated in Fig. 1. While the device appears simple enough, the actual equivalent circuit is fairly complex, as detailed in Fig. 2. The configuration applies only to the steady state. For transient conditions, e.g., sparkover and switching, the circuit is more complicated and includes the distributed parameters of the cable, precipitator, transformer windings, and rectifiers.

In Fig. 2, transformer and control circuit are shown in equivalent form, referred to the transformer secondary.[1] Not all the parameters shown in the equivalent circuit of this drawing are important in a typical precipitator installation. Values encountered in the three parts—the transformer and control circuit, rectifier circuit, and precipitator circuit—therefore invite discussion.

Transformer and Control Circuit

Tabulated herewith are typical values of transformer and control-circuit parameters, all of which are referred to the secondary of the transformer.

$L_{tm} \cong 10^4$ h (henrys)
$L_{l1} + L_{l2} \cong 10$ h
$R_{tm} \cong 10^7$ ohms
$R_s \cong 10^4$ ohms
$R_{l1} \cong R_{l2} \cong 500$ ohms

Leakage inductance L_{tm} and core-loss equivalent resistance R_{tm} are large enough to be neglected, compared with series resistances and inductances. Thus, L_{tm} and R_{tm} may be removed from the equivalent circuit and the series inductances and resistances combined to form a single series circuit of one resistance and one inductance for the transformer and control circuit. Where nonlinear controls such as saturable reactors[2] are used in the primary of the transformer, the equivalent circuit becomes more complicated and cannot be represented by a simple series combination of fixed resistance and inductance.

Rectifiers

Until well into the 1940's, synchronous rotary mechanical rectifiers were used almost exclusively with commercial precipitators. High-vacuum rectifiers became increasingly important from 1945 to 1955, and gradually replaced most mechanical types. Selenium rectifiers enjoyed a brief popularity during the early 1950's, but both they and the vacuum-tube type now are being displaced rapidly by silicon diodes.

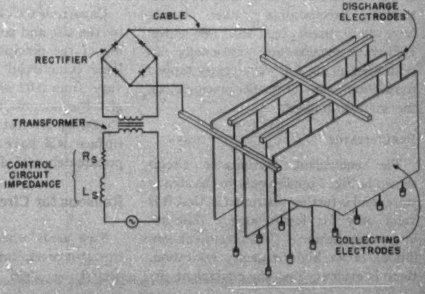


Fig. 1. An electrostatic precipitator

Paper 61-92, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 28, 1960; made available for printing November 29, 1960.

J. B. Thomas is with Princeton University, Princeton, N. J., and J. W. Drenning and H. T. Williams are with Koppers Company, Inc., Baltimore, Md.

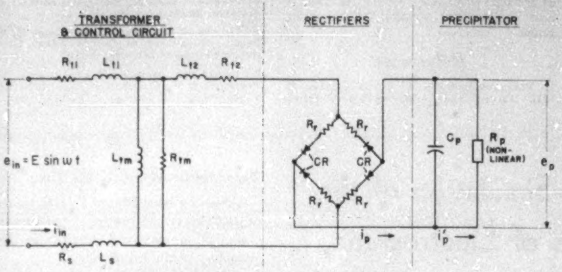


Fig. 2. Equivalent circuit of electrostatic precipitator

Each type mentioned had certain inherent advantages and disadvantages, but this paper will deal only with the two most commonly encountered today, the vacuum-tube and the silicon.

The forward resistance R_r in the vacuum-tube type is a function of the current passing through the tube, and may be determined from the rectifier characteristic curve. For typical tubes in electrostatic precipitation service, the average value of R_r is about 500 ohms and varies from 400 to 900 ohms. For silicon rectifiers, R_r is lower, a typical value for a stack falling at about 100 ohms.

Back resistance of both vacuum-tube and silicon rectifier is large enough to be ignored in this treatment. The latter unit consists of many individual diodes connected in series. Both capacitive and resistive padding of this stack may be employed for causing equal inverse voltage division to occur across each diode. These padding resistances and capacitances may be ignored in the steady-state analysis considered here, but they become important for transient conditions.

The rectifier configuration shown in Fig. 2 is the full-wave bridge circuit. This circuit and its modification, the open bridge or biphase circuit, is the one most commonly encountered in industrial electrostatic precipitators. Other circuits have been used, but less frequently. Half-wave circuits are occasionally encountered as are full-wave center-tapped connections, and, for smaller precipitators, the voltage doubler.

PRECIPITATOR

The equivalent precipitator circuit shown in Fig. 2 applies only to the steady state, and is further restricted in that it is valid only for low-frequency (line-frequency) excitation. Under transient conditions or for high-frequency excitation, there is evidence that the equivalent circuit is much more complicated.[3,4] For the precipitator alone, it consists of a capacitance C_p in parallel with a nonlinear resistor R_p, representing the corona resistance, as shown in Fig. 2.

CORONA RESISTANCE R_p

The current-voltage relationship in a given negative corona field is such that, for voltages above corona onset, the current is[5]

$$i = c_1 e^2 + c_2 e + c_3 \tag{1}$$

where c_1, c_2, and c_3 are constants. Fig. 3 shows a typical corona characteristic. The data points were measured experimentally, and the curve of equation 1 was fitted to the points with the least mean-squared error. The resulting equation is

$$i = 0.2506 e^2 - 10.58 e + 105.2$$
$$e \overset{=}{>} 26.2 \; kv \tag{2}$$

where the current i is in milliamperes (ma) and the voltage e in kilovolts (kv). Corona resistance R_p, represented by the ratio of voltage to current at each point on the curve, is shown to be quite linear. In the operating range of a precipitator, this resistance is typically in the order of 50 to 500 kilohms.

PRECIPITATOR CAPACITANCE C_p

Capacitance C_p is determined primarily by the size and collecting electrode spacing of the precipitator in question. In large commercial installations, C_p may vary from 0.02 μf (microfarad) to 0.10 μf. For example, a commercial unit with 10,000 square feet of collecting electrode surface will have a capacitance of approximately 0.05 μf.

Reasons for Circuit Simulation

Two areas where simulation of electrostatic precipitator circuits can be expected to yield valuable results are in treatment of existing installations and in development of new equipment.

TREATMENT OF EXISTING INSTALLATIONS

The modern industrial precipitator represents a challenge in testing, evaluating, and trouble shooting. An installation is inflexible to a high degree, and major components in the equivalent circuit of Fig. 2 are not easily changed. True, the series resistance R_s and series inductance L_s can be adjusted with relative ease. In addition, C_p can be increased without prohibitive difficulty.

On the other hand, capacitance cannot be decreased below the irreducible minimum inherent in the installation. Similarly, to vary corona resistance R_p involves altering wire size, wire spacing, plate spacing, or plate configuration. Any such changes tend to be of major magnitude and attainable only by overcoming many difficulties and incurring significant expenses. A similar statement can be made as to rectifiers and the transformer. To avoid such problems, the equivalent circuit can be simulated so that wide ranges of values can be substituted at will for any element

DEVELOPMENT OF NEW EQUIPMENT

Laboratory use of actual components for development purposes is expensive, cumbersome, inflexible, and frequently impossible. Likewise, existing installations are not well suited for this work. Simulation, on the other hand, provides a practical and reliable development tool with none of these disadvantages.

Simulation

Most of the circuit of Fig. 2 is easy to simulate. All elements except rectifiers

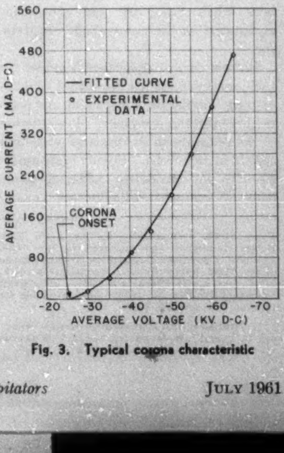


Fig. 3. Typical corona characteristic

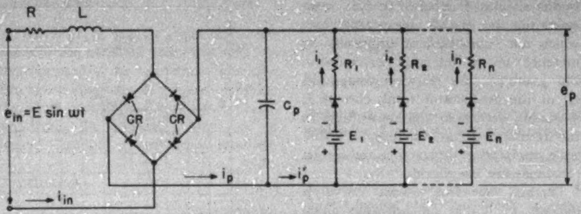


Fig. 4. Simulated equivalent circuit of precipitator

and corona resistance are lumped linear elements, obtainable in an almost unlimited range of fixed or variable values. The rectifiers can be approximated quite accurately with modern silicon diodes. Thus, there are only two serious problems in the design of an analog equivalent circuit: (1) scaling and (2) approximating the corona resistance R_p.

SCALING

This is a practical matter of selecting scaling factors which will yield component values in the analog circuit that are in the range of easily obtainable equipment. For example, such considerations make it desirable to employ potentials of less than 100 volts and capacitance values of less than 100 μf.

APPROXIMATING R_p

The corona characteristic in Fig. 3 can be divided conveniently into two parts: voltages less than and those above corona onset. The characteristic of zero current is easily obtained with biased diodes in the first case, and a number of possibilities for approximation exist in the second. Generation of the curve by using a combination of linear and nonlinear resistors is an obvious approach. The latter are readily available in the form of thyrite rods and disks. Enough work has been done in this area of approximating nonlinear current-voltage relationships,[6] so that a close fit can be obtained. However, the resulting circuit is relatively inflexible, may be fairly complicated, and requires careful adjustment of the nonlinear resistors This adjustment may be impractical since only a limited number of nonlinear resistance values are readily available.

Another possibility is to approximate the corona curve by a series of straight-line segments, each segment being generated by a linear resistor and a biased diode.[7] This approach has the advantage of simplicity, although a large number of segments may be required if highly accurate reproduction is desired. As will be shown later, sufficiently accurate approximations to typical corona curves can be obtained with relatively few segments. Because of the resulting simplicity of the circuit, this approach will be treated here.

A Practical Realization

The schematic diagram, Fig. 4, shows a practical method for simulating typical equivalent circuits of industrial precipitators. The magnetizing and core-loss components of the transformer circuit have been neglected, and all series resistances and inductances in the transformer, control circuit, and rectifier have been lumped into a single-series combination denoted by R and L. The rectifiers are low-leakage silicon diodes, and can be treated as ideal. A linear capacitor C_p represents the precipitator capacitance. The corona resistance R_p is approximated by a set of n parallel circuits, each consisting of a linear resistor and a biased diode in series combination. Each parallel circuit thus provides one straight-line segment for the corona-curve approximation.

The usefulness of the circuit just described depends largely on the value of the integer n. Experience indicates that, for most corona curves encountered in practice, this number need be no greater than four. In fact, it will be shown that $n=2$ can accurately approxi-

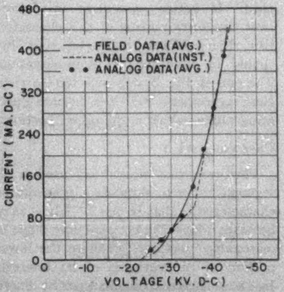


Fig. 5. Comparison of actual and simulated corona characteristics

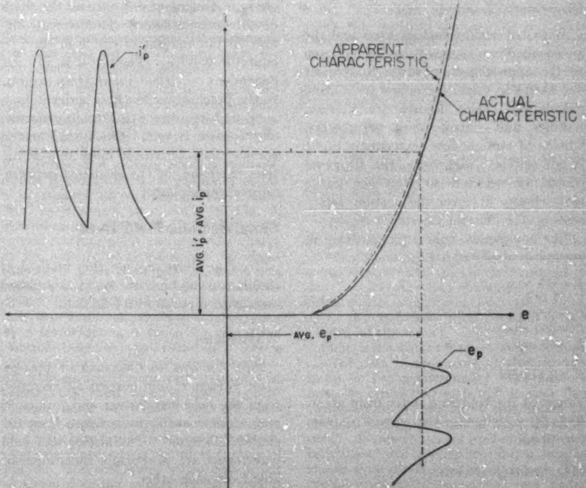


Fig. 6. Instantaneous current-voltage relationships in precipitator

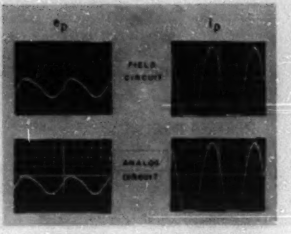


Fig. 7. Comparison of waveforms from actual and simulated precipitator circuits

mate the corona curve representative of many industrial installations. This curve is shown in Fig. 5; the solid line was obtained from an actual precipitator installation and the two dashed-line segments are an approximation corresponding to the case $n=2$. This example will now be considered in detail.

SCALE FACTORS

It is convenient to select scaling factors which permit the use of standard laboratory equipment for the simulated circuit. Since power supplies are most readily available in the range 0 to 500 volts and 0 to 100 ma, a voltage scaling factor of 0.002 and a current scaling factor of 0.10 are found to reduce the full-scale values to the proper ranges for the analog circuit.

COMPONENT VALUES

With the scaling factors fixed and the two dashed-line segments in Fig. 5 chosen as the approximating curve, component values in Fig. 4 are determined by routine calculations, using known transformer, rectifier, and control-circuit parameters. Choice of straight-line segments in Fig. 5 is not critical. Generally, the segments should be selected so that the curve most closely fits over the portion representing the normal operating region.

The component values for the case in point are calculated to be:

$R = 617$ ohms
$L = 0.48$ h
$C_p = 2.50$ μf
$R_2 = 630$ ohms
$E_1 = 46$ volts
$E_2 = 71$ volts
$R_1 = 2{,}412$ ohms

Values of the voltage e_p vary from about 40 to 90 volts and of the current i_p from 0 to 40 ma.

MEASURED QUANTITIES

Fig. 4 demonstrates that all important currents and voltages can be measured in the simulated circuit. In fact, some quantities are readily measurable here which are impractical or impossible to measure on a full-scale precipitator. For example, the resistive component i_p' of the precipitator total current i_p is readily accessible and measurable in the simulated circuit, whereas, on a full-scale precipitator, capacitance and corona resistance are inseparable.

Further, the true corona curve is difficult to obtain on full-scale units. Measurement of average current and average voltage will not yield the true curve as illustrated in Fig. 6. A typical actual characteristic, i.e., the instantaneous characteristic, is shown by a solid curve in that figure. This is the curve that would be obtained through the use of flat d-c excitation. A typical precipitator voltage waveform e_p is also shown, and, by projecting it point-by-point on to the instantaneous corona curve, the corresponding instantaneous resistive component i_p' of precipitator current is obtained. The average value of i_p' will be equal to the average value of i_p since the average steady-state current through C_p is zero.

Average precipitator voltage and current do not intersect the actual corona curve coincidentally when they are projected. Thus, plotting a corona curve from average-reading d-c meters will yield an apparent curve, which exhibits a slope greater than that of the instantaneous curve. Moreover, this apparent curve is dependent not only on the shape of the instantaneous characteristic, but also on the amount of ripple in the precipitator voltage. The latter is, in turn, dependent on circuit parameters external to the precipitator, such as series inductance and resistance in the transformer and control circuit. The instantaneous corona characteristic is more representative, therefore, of the electrical performance of a precipitator.

Simulated and Field Data

To show how reliable this method of simulation can be, some typical simulated data are compared with field data.

WAVEFORMS

Fig. 7 shows oscillograms of precipitator voltage e_p and current i_p, recorded from the field installation and compared with similar oscillograms taken from the simulated circuit. The analog and field waveforms are practically indistinguishable from each other.

Fig. 8 shows the current waveform reproducibility over a range of operating voltages. The ratio of average to effective current is a measure of the ripple in the waveform, and is a convenient parameter for quantitative comparison. This ratio has been plotted in the drawing as a function of average operating voltage, and data points from the corresponding field installation are given by way of comparison. Deviation between field and simulated data is small enough to be unimportant for engineering purposes; that which does exist is due primarily to difficulty in obtaining accurate field measurements.

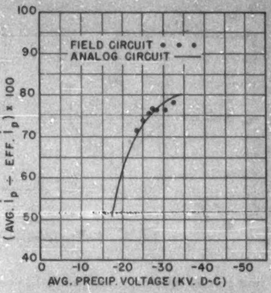


Fig. 8. Comparison of current form-factors from actual and simulated precipitator circuits

CORONA CHARACTERISTIC

The reproducibility of the corona characteristic is shown in Fig. 5, where departure of the average analog data points from the 2-segment instantaneous curve is apparent. However, the versatility of the simulated circuit allows parameters to be readily adjusted so as to fit the average analog points to the average field data curve. Over the normal operating range of precipitator voltage, reproduction of the corona characteristic is within the accuracy of the instruments used. If better agreement is desired over the lower portion of this characteristic, then a different pair of line segments should be used for the approximation. The pair in Fig. 5 was selected to yield close agreement over the normal operating range, which is the upper portion.

CASE FOR $n=1$

The case just illustrated used a 2-segment curve to simulate a precipitator characteristic, that is, $n=2$. There may be cases where the corona-curve slope is reasonably constant or where the lower portion of the curve is of no interest. In these cases, a single-segment curve ($n=1$) will give surprisingly accurate results.

Case for $n>2$

Values of n greater than two will normally be required only where it becomes essential to reproduce faithfully the entire characteristic curve or where the characteristic has a relatively large curvature. In general, values of n greater than four will seldom be required.

Conclusions

A relatively simple equivalent circuit describes accurately the electrical characteristics of commercial precipitator installations. With appropriate scaling, this circuit may be simulated readily and inexpensively with conventional laboratory equipment. The resulting analog is highly flexible, and can be adjusted to simulate changes in precipitator operating conditions and equipment.

The use of analog techniques is desirable and necessary in fields such as electrostatic precipitation where the typical installation is large, inflexible, and complex. These techniques can contribute significantly to design, development, and testing in this field.

Nomenclature

Definitions are given herewith for terms used in Fig. 2, and elsewhere in this paper.

R_s = series resistance of control circuit referred to transformer secondary
L_s = series inductance as above
R_{t1} = resistance of transformer primary winding referred to transformer secondary
L_{t1} = inductance of above winding
L_{tm} = transformer leakage inductance referred to transformer secondary
R_{tm} = transformer core-loss equivalent resistance referred to transformer secondary
R_{t2} = resistance of transformer secondary winding
L_{t2} = inductance of above winding
R_r = rectifier forward resistance
CR = ideal rectifier
C_p = equivalent capacitance of precipitator
R_p = equivalent resistance of precipitator
i_{in} = instantaneous input current
e_{in} = instantaneous input voltage
i_p = instantaneous precipitator current
i_p' = instantaneous resistive component of precipitator current
e_p = instantaneous precipitator voltage

References

1. Transformer Engineering (book), L. F. Blume, editor. John Wiley & Sons, Inc., New York, N. Y., 1943.

2. Saturable Reactor Control of Full-Wave and Biphase Rectifiers, J. B. Thomas, J. W. Drenning. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, Mar. 1960, pp. 36–39.

3. A Pulse Method for Supplying High-Voltage Power for Electrostatic Precipitation, H. J. White. *Ibid.*, vol. 71, Nov. 1952, pp. 326–30.

4. Current-Voltage Relationships in Negative Pulsed Corona, J. B. Thomas, T. R. Williams. *Ibid.*, vol. 79, May 1960, pp. 136–39.

5. Gaseous Conductors (book), J. D. Cobine. McGraw-Hill Book Company, Inc., New York, N. Y., 1941, chap. 8.

6. The Generation of Squares with the Use of Non-linear Resistors, E. Grosswald. *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-6, no. 4, Dec. 1959, pp. 334–39.

7. Non-linear Electrical Networks (book), W. L. Hughes. The Ronald Press, New York, N. Y., 1960, chap. 2.

◆

Discussion

Gaylord W. Penney (Carnegie Institute of Technology, Pittsburgh, Pa.): The scaling down of various components, which will permit experimental work at low voltages and therefore with small laboratory apparatus, is quite attractive to me. However, even though constants are adjusted so as to produce an accurate replica of curve, I still regard the analog as a mere approximation, and am inclined to question the first sentence in the conclusions. There the authors state that the "equivalent circuit describes accurately the electrical characteristics of commercial precipitator installations."

For making calculations, I have often used the circuit shown in Fig. 4, and, even though the d-c current-voltage curve was duplicated exactly, I still regarded the equivalent circuit as an approximation.

The energy storage in a corona discharge is so different from that in a conventional capacitor that obtaining the proper value of capacity seems questionable. The energy storage hardly can be represented by any fixed capacity. The usual definition of capacity is charge divided by voltage, which applies when the charges are on fixed electrodes with a material of constant permittivity between the electrodes. Attempts to apply this definition to corona reveal that charge is not proportional to potential difference. A paper written by me in collaboration with Matick[1] gives curves of potential versus position for several typical conditions. From these curves, the potential gradient at the electrodes can be obtained, and the charge on the electrodes calculated.

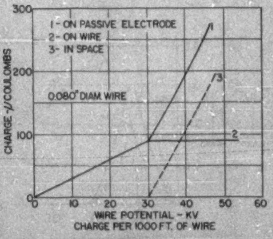


Fig. 9. Charge on electrode surfaces as computed from measurements of space potentials in d-c corona

Figs. 16 and 17 in that paper give potentials for wire size and electrode spacings that are quite typical of industrial precipitators, and Fig. 9 shows plots of charge versus voltage, corresponding to the data in 16 and 17. Below the corona starting voltage, charge is proportional to voltage, and the usual concept of capacity applies. However, as is shown by curve 1 of Fig. 9, increasing the voltage to 50% above the corona starting point increases the charge on the large or passive electrode by about threefold. Thus, if using the definition that capacity is charge on the passive electrode divided by applied voltage, a 50% increase in voltage would raise the calculated capacity by a factor of about two. On the active electrode behavior is quite different. Above the corona starting potential, the charge does not change appreciably with voltage. Of course, the total positive and negative charges must be equal. The difference between charges on passive and active electrodes represents the charge in the interelectrode space, and is shown as curve 3 of Fig. 9. The best value of capacity in this situation does not appear obvious, and it seems doubtful if any particular capacity can be used over a wide range of conditions.

Because of this uncertainty and to help the reader in applying this equivalent circuit to other situations, would the authors specify the wire size and other essential dimensions corresponding to the value of capacity in the example given under the heading "Component Values?" Also, what was the condition of the wire—new, etched, clean, or with accumulated dust? Also, referring again to the first sentence in the conclusions, over what range of conditions does the statement "equivalent circuit describes accurately the electrical characteristics of commercial precipitator installations" apply? For example, could the circuit be used for a 400-cycle power supply and, if so, would the same value of capacity be used for both 400 cycles and for 60 cycles?

Reference

1. Potentials in D-C Corona Fields, Gaylord W. Penney, Richard E. Matick. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, May 1960, pp. 91–99.

P. L. Smith (Carnegie Institute of Technology, Pittsburgh, Pa.): The authors have contributed a useful idea in their concept of deriving an equivalent circuit for study of electrostatic precipitator behavior. However, two questions arise concerning the validity of this circuit.

First, is the instantaneous voltage-current characteristic of the corona identical to the static or d-c characteristic? The authors state, in connection with Fig. 6, that this characteristic is "the curve that would be obtained through the use of flat d-c excitation." However, as in conventional electron tubes,[1] this remains true only so long as transit time effects may be neglected. For typical precipitator dimensions and under typical operating conditions, the transit time of an ion across the interelectrode gap is of the order of one millisecond. Thus, for changes requiring time which is long compared to one millisecond, the authors' assumption is reasonable. But even at 60 cycles, with full-wave rectification, the period of the precipitator voltage waveform is one-half cycle, or about 8.3 milliseconds. Thus, 60 cycles is close to the upper limit of the frequencies range for which this assumption is valid.

The second question, concerned with determining the value of capacitance C_p, arises here because part of the precipitator charge is distributed through the interelectrode space, rather than all of it being located on fixed electrodes. Thus the precipitator capacitance is difficult to determine, if, in fact, a fixed value of capacitance can be specified.

The total charge (of one sign) in a precipitator increases faster than the applied voltage after the onset of corona; consequently, if one uses the definition

$$C=\frac{Q}{V}$$

to calculate the capacitance, the conclusion is that capacitance must be a function of voltage. Even if the form

$$C=\frac{dQ}{dV}$$

is used, it has not been demonstrated that a constant value of capacitance will result, or that the value would be the same as the ordinary low-voltage capacitance of the precipitator. This question must be more carefully examined if the quantitative validity of the proposed equivalent circuit is to be established.

It also appears to the writer that the time scale on the oscillograms of Fig. 7 has been reversed; that time increases from right to left. Is this the case?

Reference

1. Radio Engineers' Handbook (book), F. E. Terman. McGraw-Hill Book Company, Inc., New York, N. Y., 1943, pp. 308–12.

J. B. Thomas, J. W. Drenning, and **H. T. Williams:** The discussions by Professors Penney and Smith pose several questions which are not specifically treated in the paper. As mentioned therein, the simulation was of large industrial precipitators of the type diagrammed in Fig. 1, and ranges of parameter values were typical of such installations. In these units, wire-to-plate dimensions are in the order of 3 to 5 inches; wire-to-wire spacing is approximately 5 to 6 inches; and wire diameters are approximately 0.1 inch. In some designs, wires do not have circular cross sections, but have corona characteristics similar to those of 0.1-inch-diameter round wire. The simulation was designed for operating installations where varying amounts of dust are present on both electrode systems and where the surface conditions of the active electrodes are extremely variable. In such installations, simple rectifier circuits as in Fig. 2 are energized from commercial power lines, the majority of which operate at 60 cps (cycles per second). Where 60 cps is not used, 50, 40, or 25 cps (i.e., lower frequency) power is available. As stated in the paper, "The equivalent precipitator circuit shown in Fig. 2 applies only to the steady state, and is further restricted in that it is valid only for low-frequency (line-frequency) excitation. Under transient conditions or for high-frequency excitation, there is evidence that the equivalent circuit is much more complicated."

The choice of a lumped linear capacitor and a parallel nonlinear resistor as the equivalent precipitator circuit was dictated and justified by the observation that such a combination reproduces the precipitator terminal behavior observed under the conditions set forth in the paper and in the previous paragraph. The points raised by both Penney and Smith concerning capacitance values should be given further attention.

A logical extension of the subject work would be to consider the simulation problem over a broad frequency range above commerical line frequencies.

Mr. Smith has pointed out that the time scale in Fig. 7 runs from right to left. This is correct, although no time scale is explicitly shown. The oscillograms are given for comparison purposes.

Switching Characteristics of Silicon Power-Controlled Rectifiers

I—Turn-On Action

I. SOMOS
MEMBER AIEE

TURN-ON TIME is the time interval required for decreasing the forward voltage occurring across a cell to 10% of its initial value after a gate current is applied. It contains two parts: (1) delay time—the period needed for the voltage to reduce from 100% to 90%, and (2) rise time—the period during which the voltage reduces from 90% to 10%.

The rise time is actually a decay time, but the first term is used because it is in accordance with the general definition of a transient.

Turn-on actions can occur in three ways:

1. Turn-on action by voltage occurs when the forward voltage exceeds the breakdown level, and the SCR (silicon controlled rectifier) turns on; then when the current flowing in the circuit is larger than the holding current, it remains in the conducting state.

2. High rate of change of the forward voltage also can cause turn-on action. When a forward voltage of steep front occurs across the SCR, the transient capacitive current which flows is determined by the middle junction, whereas the side junctions are forward biased. When the rate of rise is sufficiently high, the capacitive current causes the SCR to fire, even though the peak value of the voltage wave is smaller than the static breakdown voltage. Fig. 1 shows the voltage and current of an SCR of 5/16-inch-diameter area, taken with a double-trace oscilloscope and coaxial shunt. The current shown in the picture caused the SCR to fire. The d-c static self-firing voltage was 110 volts, and the d-c static gate current was 140 ma (milliamperes). The line showing the current i_{cap} was drawn by a pure capacitor, subjected to the same voltage.

The capacity of the SCR is a function of the voltage. The capacity of the middle junction can be calculated from the curves in Fig. 1 by the relation

$$C=\frac{i}{dV/dt}$$

for voltage values between 0 and 40 volts. In this way, curves of C as a function of voltage can be obtained.

Fig. 2 shows the border line of the self-firing area for room temperature and for 60 cps (cycles per second) for an SCR, having 280-volt static breakdown voltage. Self-firing

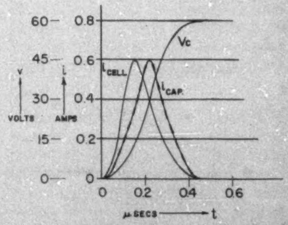


Fig. 1. Capacitive current flowing through SCR and through a capacitor

---

Paper 60-227, recommended by the AIEE Semiconductor Metallic Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted November 30, 1959; made available for printing December 2, 1960.

I. Somos is with General Electric Company, Collingdale, Pa.

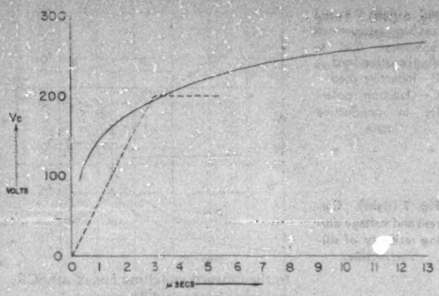
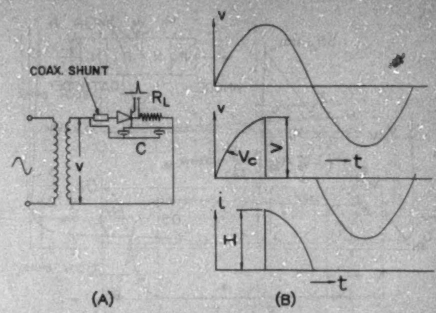


Fig. 2. Critical self-firing curve



(A) (B)

Fig. 3. Circuitry used for testing turn-on action (A), and voltage-current waves (B)

occurs when the wave front of the applied square voltage intersects the curve, as shown by the dashed line. The frequency influences the self-firing point because, in the OFF time of the square wave, the middle-junction capacitor discharges and, in the following cycle, this junction will be hit in partly charged condition. The amount of discharging in the OFF time depends on the frequency. The total discharging of the junction capacitor lasts a relatively long time. It may be 0.1 second or more. At more than 1,000-cps frequency, the voltage curve across the middle junction approaches a straight line.

3. Gate current turn-on action is the form described in this paper. All test results were obtained on 5/16-inch–diameter silicon controlled rectifier.

Interpretation by Firing Diagram

The turn-on process may be visualized from examination of voltage versus current, as diagrammed during the switching transient and with the circuits as shown in Fig. 3(A). The SCR was fired at 90 degrees; in other words, just at the top of the sinusoidal voltage wave. Fig. 3(B) shows the cell voltage V_c and current.

In the circuit, Fig. 3(A), the effect of the inductance was eliminated by using different types and sizes of capacitors. A combination of paper, Pyranol (a carefully purified, chlorinated hydrocarbon), and electrolytic capacitors effectively cancels the inductance of the power supply over a frequency range in direct current from 0 to 10 mc/s (megacycles per second).

Symbols V and I indicate, respectively, the instantaneous value of the voltage prior to firing, and the current after firing. Since the turn-on action represents a minimal part of the 60-cps wave, the turn-on process is simiar to that observed in a d-c circuit.

The circuits in Fig. 3(A) were used for most of this report. Firing diagrams, Figs. 4 and 5, were obtained from these circuits. The negative values are not drawn.

Fig. 5 shows the lower part of the curve with an enlarged current scale. When no gate current is flowing, the saturation current equals $I_s^{(o)}$. If gate current is applied, saturation current will be increased nearly proportional to gate current. When the gate current reaches a certain limit, the SCR fires. This value is 80 ma, as shown. Saturation current resulting from this gate current is I_s'. For any change of gate current between 0 and 80 ma, a transistor-like effect is present, saturation current following the form of the gate current.

When gate current exceeds the limit value necessary for firing and the SCR turns on, the operating point proceeds to the steady-state point, as determined by the forward characteristics of the SCR (dotted line in Fig. 4) and by the load.

The firing characteristic, shown in Fig. 4, is for a pure resistive circuit A, and for an inductive circuit B. For both cases, the time diagrams of cell voltage and current are given also. The form of the firing diagram is independent of turn-on time and depends only on the circuit. The pure resistive load is indicated by a straight line, and inductive circuits by a curve.

If two out of three diagrams (firing, cell voltage versus time, and current versus time) are given, the third can be obtained. Point P of the voltage diagram in Fig. 4(A) is obtained from points P_1 and P_2.

The time points on the firing diagrams

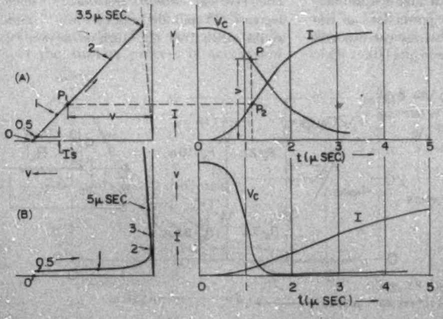


Fig. 4 (left). Firing diagrams

A—Resistive load
B—Inductive load

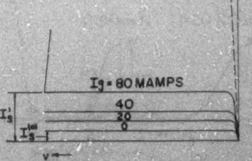


Fig. 5 (right). Saturation-current curves for various gate currents

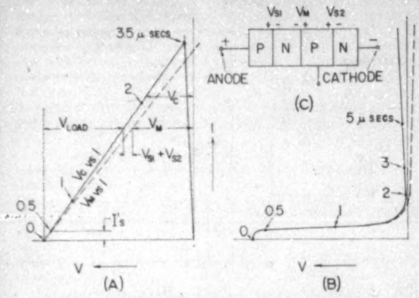


Fig. 6 (left). Firing diagrams

A—Resistive load
B—Inductive load
C—Junction polarity in conductive state

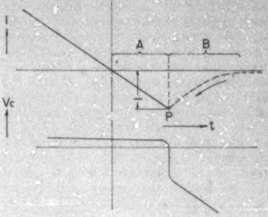


Fig. 7 (right). Current and voltage during recovery of silicon rectifier

indicate the firing-action procedure. The first part of the diagram—up to 0.5 μsec (microseconds)—indicates the time interval during which the saturation current increases from $I_s^{(o)}$ to I_s'. This interval, which can vary between wide limits, depending on the gate-current form and size and the load, constitutes the delay time, during which the cell voltage changes only slightly. Gate current of a rise time much shorter than the delay time itself must be applied in measuring the latter.

Inversion of Voltage Across Middle Junction

Figs. 6(A) and 6(B) show V_m, the voltage across the middle junction during turn-on action. The solid and dashed lines indicate the V_c versus I, and V_m versus I, respectively. The difference between V_c and V_m is equal to the voltage drop across the two side junctions ($V_s = V_{s1} + V_{s2}$). Separate voltage drops cannot be measured exactly, but V_s can be determined from the conventional forward characteristic, with consideration given to the areas of the side junctions.

Solid line V_c versus I can be observed on the oscilloscope. By calculating V_s, the dashed line can be obtained. This line intercepts the Y-axis, and the portion that crosses the Y-axis shows the polarity inversion.

In the TURNED-ON condition, voltage distribution across the cell corresponds to Fig. 6(C). The middle junction is forward biased, while the current I flowing through it is a reverse current with respect to this junction. The middle junction in this condition is, therefore, similar to the photocell or to the recovery state of a silicon rectifier. In this state, the rectifier can be compared to a generator; the vectors of its forward voltage drop, and the current flowing through it have opposite polarities. The product $V \times I$ represents the output of the generators, causing no heat in the rectifier. A similar condition is found in the middle junction of the SCR.

The similarity can be understood from Fig. 7, which shows the commutation of the silicon rectifier in a single-phase circuit. Before firing the SCR, the central junction is reverse biased, corresponding to B, and firing can be compared to a reverse procedure, shown by the arrows. The dashed line corresponds to similarly broken lines in Fig. 6. The current increases to point P, and there the polarity of the voltage changes. The same change occurs in the middle junction of the SCR, as may be visualized from the turn-on curves in Figs. 8(A) and 8(B), resistive and inductive loads.

The transient power-loss circumstances are similar for the silicon rectifier and SCR. In Fig. 7(B), the current and voltage have the same polarity, and heat is developing, indicating a high-instantaneous but small-average power loss of the rectifier. This power loss corresponds to the forward switching loss of an SCR, occurring in the middle junction.

There is, however, a difference between them. The generator-like behavior of the rectifier is a short cleaning-out transient. The generator characteristics of the middle junction of the SCR is maintained continuously by the side junction. Thus, the middle junction could be compared to a converter having input and output energy. This situation arises when the forward watts loss in the SCR is calculated.

The forward power loss is less than the sum of the theoretical power losses of the two side junctions, and may be calculated from the actual voltage drop and current as in simple rectifiers.

Influence of Resistive Circuit

The turn-on process was defined by the time diagram of the cell voltage, where the curve is influenced by the load. Consider the SCR as an ohmic resistor $R_c^{(t)}$, which changes its resistance during the turn-on action. The voltage V will be distributed between R_L and $R_c^{(t)}$ during the transient, as shown in Fig. 9. $R_c^{(t)}$ indicates the instantaneous values of the cell resistance.

Fig. 10 shows the instantaneous values of $R_c^{(t)}$ calculated from the turn-on curves, for 4 amperes and 20 amperes at 40 volts. The rate at which resistance values decrease is rapid for high currents. This would mean that the turn-on action is,

Fig. 8 (left). Turn on curves

A—Resistive load
B—Inductive load

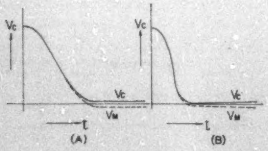
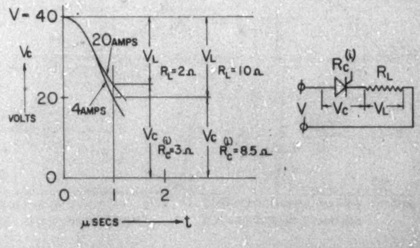


Fig. 9 (right). Transient voltage distribution between SCR cell and load during turn-on action



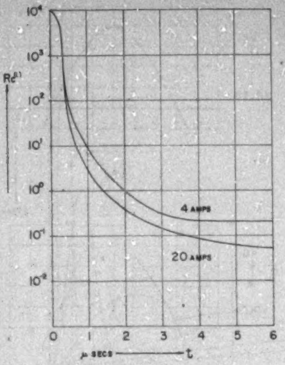
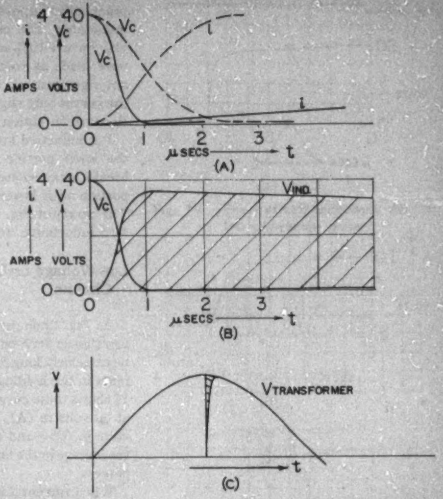


Fig. 10 (left). Instantaneous values of SCR resistance during turn-on action

Fig. 11 (right). Turn-on curves for resistive and inductive load (A), voltage across inductance (B), and voltage across transformer (C)



in effect, faster for higher currents. However, the turn-on curve will be determined by

$$V_c = V \frac{R_c^{(t)}}{R_L + R_c^{(t)}}$$

Although R_c^t is smaller for higher currents at any time point during the transient,

$$\frac{R_c^{(t)}}{R_L + R_c^{(t)}}$$

is larger. This means that, for a given time point, the cell voltage V_c is higher for high currents (see Fig. 9 for V equal to 40 volts and I equal to 4 amperes and 20 amperes, the given values corresponding to 1 μsec). In other words, the turn-on curve is longer for higher currents.

Influence of Circuit Inductance

If inductance is present in the circuit, the voltage (V in Fig. 9) will be distributed between the cell, load resistance, and the reactance during the transient. In many cases, the reactance may be so high that it represents the highest impedance of the circuit during the transient. The cell voltage, therefore, decreases rapidly and the turn-on process is accelerated. In other words, the reactance absorbs the voltage from the SCR, and voltage across the middle junction breaks down in a short time. Current increases slowly, its rate of rise being controlled by the time constant of the circuit L-R (inductance resistance), instead of the semiconductor device.

In Fig. 11(A), voltage and current are given for a resistive load (dashed line) and for an inductive load. The initial slope of an exponentially rising current is

$$\left.\frac{dI}{dt}\right|_{t=0} = \frac{I}{L\text{-}R} = \frac{V}{L}$$

This means that the initial current slope is independent of the current I and load resistance. Hence, for a given inductance L and a voltage V, the same turn-on curves will be obtained over a wide current range. The inductive turn-on curve in Fig. 11(B) for $L=0.25$ mh (millihenry) does not change substantially for a range of $I=1$–100 amperes.

The voltage absorbed by the inductance is

$$V_{\text{ind}} = -\frac{dI}{dt} L$$

As shown in Fig. 11(B), the overwhelming part of the voltage occurs transient-wise across the inductance. In conventional circuits, inductance is represented mainly by the transformer. Thus, the secondary of the transformer will be modified as shown in Fig. 11(C). The shaded areas in (B) and (C) are identical.

The values in Fig. 11 are: $V=40$ volts, $I=4$ amps, $R_L=10$ ohms, $L=0.25$ mh. The calculated current slope at $t=0$ for the exponential current curve is: $dI/dt=$ 0.16 ampere/μsec. The measured value at $t=1$ μsec is: $dI/dt=0.14$ ampere/μsec. Voltage across the inductance at $t=1$ μsec is $V_{\text{ind}}=0.25\times10^{-3}\times0.14\times10^{6}=35$ volts.

As was shown, inductance decreases the turn-on time. However, when inductance exceeds a certain limit, its effect will be

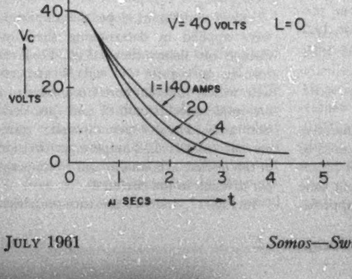


Fig. 12 (left). Turn-on curves for resistive load

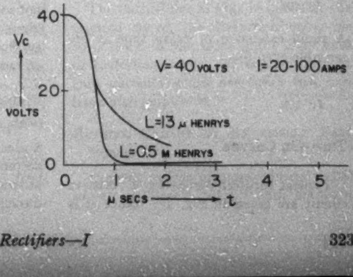


Fig. 13 (right). Turn-on curves for inductive load

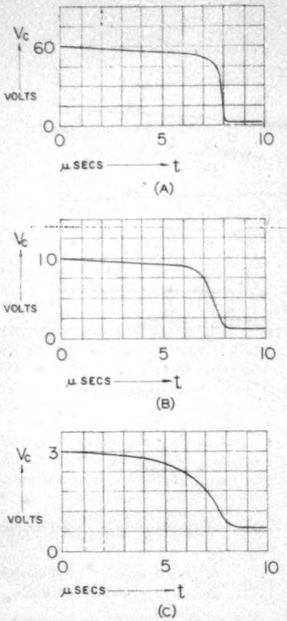


Fig. 14. Turn-on curves for low current I = 25 ma

the opposite, and turn-on time increases. In this case, the time to reach the firing current I_s' may be long, and an effect similar to slow switching will take place.

As known, the current slope equals V/L. For a current slope, which permits reaching I_s' in 1.0 μsec, V/L must be larger than

$$\frac{I_s'}{10^{-6}}$$

It follows that, in order to avoid slow switching or prolongation of delay time, the inductance must be smaller than

$$L=\frac{V}{I_s'\,10^{6}}$$

A small extension of delay time is observable even for

$$L=\frac{V}{I_s'\,10^{7}}$$

Turn-On Curves

Typical turn-on curves for a resistive circuit are given in Fig. 12. Since it is practically impossible to eliminate all inductance, the curve for 140 amperes may have +10% error. While the curves were made at room temperature, this is not significant as temperature influences the curves only slightly. Above 40 volts, the turn-on action is somewhat faster.

For inductive loads, see Fig. 13, where the lower portion of the turn-on curve breaks away sooner from the steep-slope portion with lowered circuit inductance. The curves of Fig. 12 were obtained when zero inductance was reached.

Low-Voltage or Low-Current Circuits

As Fig. 12 shows, lower current results in a shorter turn-on curve, although it will be extremely long for extremely small currents in the holding-current region. Fig. 14 shows these curves for 25-ma current—at 60 volts in (A), 10 volts in (B), and 3 volts in (C)—and all depict the effect of the voltage in the latter part of the turn-on process.

When current I is lower than the holding current, it flows only for the duration of the gate-current pulse. Fig. 15(A) represents a gate-current pulse of 300-μsec duration, and shows the cell voltage versus time.

In Fig. 15(B), current I is approximately equal to the holding current. For the duration of the gate-current pulse, the instantaneous value of the voltage is sufficiently high so that I is higher than the holding current. The turn-on process will, therefore, be accomplished. In the latter part of the cycle when voltage decreases, I goes below the holding current and the cell stops conducting.

Size and Form of Gate-Current Pulse

The preceding analysis was made with a gate-current pulse of approximately 300-μsec length and 30-μsec rise time. The amplitude of the pulse was approximately double the minimum necessary d-c static value. The rise-time values may be close to those used in many practical applications; a small deviation does not mean any change in data given previously. However, when fast or high gate-current pulses are applied, the turn-on curve changes substantially.

FIRING WITH SHORT PULSES

When pulse duration is shorter than 6 to 8 μsec, a larger amplitude is required to accomplish firing. For a series of pulses having the same shape but different time durations, as in Fig. 16, there is approximately an inverse correlation between pulse length and amplitude, over the range of 0.2- to 2-μsec pulse lengths.

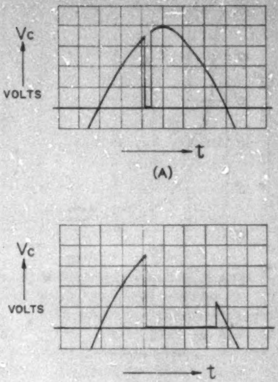


Fig. 15. Turn-on action for load current approximately equal to holding current

The shortest pulse in this figure had a total length of approximately 0.18 μsec. A length of 0.1 μsec was not adequate for firing, even at 7 amperes amplitude. There seems to be a practical limit in the vicinity of 0.1-μsec pulse length, under which firing is not possible. The curves shown were obtained at 50 volts, 10 amperes output load at room temperature. The d-c gate current was 80 milliamperes.

When pulses of such short duration are applied, the inductance of the wiring between pulse generator and gate terminals must be low. A slight change in inductance may greatly alter the pulse form, and the SCR may even be switched ON and OFF by a barely perceptible variation in gate-current inductance. In this case, increased inductance caused the SCR to switch ON, and decrease caused it to turn OFF.

GATE-CURRENT PULSES AND DELAY TIME

Pulses of different slope and amplitude were applied in determining their influence on delay time. Fig. 17 gives turn-on curves for 50 volts/4 amperes with widely different currents applied in a low-inductance circuit (1 μsec per division and 12 volts per division; gate-current scale is 0.18 ampere per division in the circled 1, 2, and 3, and 10 amperes per division in the circled 4).

In Fig. 17, 1 shows the turn-on action

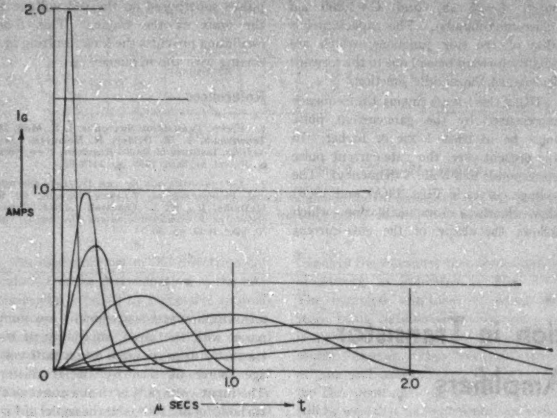


Fig. 16. Gate-current pulses just sufficient to fire

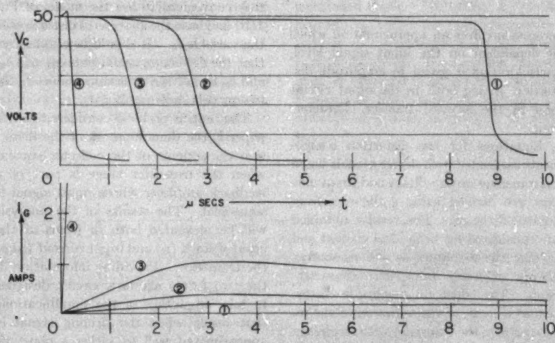


Fig. 17. Delay time for different gate currents

for a gate current, which is barely able to fire because of slow switching for which the delay time is extremely long—here, 9 μsec. When the gate pulse increased, as in 2 and 3, the delay time was shorter. When pulses attained quite high initial slopes such as 10 amperes per μsec, delay time diminished to approximately 0.1 μsec.

The SCR delay time may be compared to the rise time of a transistor. If gate currents are too small to fire, the SCR conforms to a common emitter transistor with collector current i_C and base current i_B corresponding to the saturation and gate currents, respectively. The current i_C does not occur immediately after application of base current. The period required to reach from 10% to 90% of its final value is indicated by rise time t_r.

As known

$$t_r=\left[\frac{1}{(1-\alpha_{fb})\omega_{\alpha fb}}\right]\ln\left[\frac{i_B}{i_B-0.9(1-\alpha_{fb})i_c/\alpha_{fb}}\right]$$

where α_{fb} is the common base-current gain, $\omega_{\alpha fb}$ is the common base-cutoff frequency. As this equation shows, for large i_B values, t_r approaches zero. Its values range substantially between 0.1 to 20 μsec, as do those of the SCR delay time.

Interpreted from the firing diagram, Fig. 4, delay time is the interval from zero to 0.5 μsec.

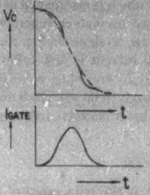


Fig. 18. Effect of cell inductance due to fast gate-current pulses

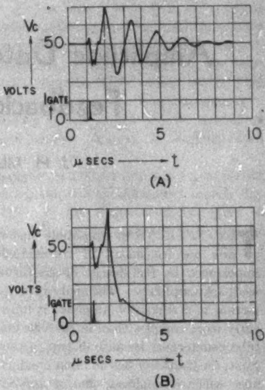


Fig. 19. Ringing caused by very fast gate-current pulses

A—Gate current unable to fire
B—Gate current just sufficient to fire

DISTORTIONS BY FAST GATE-CURRENT PULSES

The large slope of the gate current—6×10^6 amperes per second—and inductance of the cell across the gate caused the distortion appearing in Fig. 18. The induced voltage $L\,dI$ gate$/dt$ is approximately 3 volts. This means the inductance of the gate-top junction is equal to $L=3/6=0.5$ μh. Other testings gave approximately the same value.

The dashed line in Fig. 18 indicates the original curve, the distortion being incorporated in the solid line. This kind of distortion occurs, of course, only at gate currents which are fairly fast and have high amplitude.

RINGING CAUSED BY EXTREMELY FAST PULSES

Current pulses of approximately 100 mμsec (milliμsec) total length create a ringing, which is manifested in the form of a damped oscillation superposed on the cell voltage, as shown in Fig. 19. In 19(A) the gate current is just below the value necessary to fire, while in 19(B) it is just high enough to do so. In both cases, the ringing has a resonant frequency

$$f_r = \frac{1}{2\pi\sqrt{LC}}$$

where L is the inductance of the circuit and C is the capacity of the cell. The value obtained for this case where $f=1$ mc/s, $L=25$ μh, and $C=1{,}000$ μμf (micromicrofarads). This capacitance is that of the side junctions, which are slightly forward biased due to the forward leakage of the middle junction.

To get this 1-mc/s ringing, the frequency represented by the gate-current pulse must be at least 1 mc or higher. In the present case, the gate-current pulse corresponds to a 5-mc/s frequency. The voltage curves in Figs. 19(A) and 19(B) show clearly a 5-mc oscillation, which follows the shape of the gate-current pulses superposed on the cell voltage at the start of the ringing. This 5-mc oscillation provides the kick resulting in a ringing with the frequency f_R.

References

1. PNPN TRANSISTOR SWITCHES, **J. L. Moll, M Tanenbaum, J. M. Goldey, N. Holonyak.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 44, Sept. 1956, pp. 1174–82.

2. AN INVESTIGATION OF THE DYNAMIC SWITCHING PROPERTIES OF 4-LAYER DIODES, **G. E. McDuffie, Jr., W. L. Chadwell.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, Mar. 1960, pp. 50–55.

Amplitude Distortion in Transistor Feedback Amplifiers

J. H. MULLIGAN, JR.
MEMBER AIEE

THE EFFECTS of amplitude distortion are frequently of considerable importance in the design of transistor feedback amplifiers. The analysis of distortion in a transistor amplifier is inherently more complex than in the vacuum-tube counterpart because of two factors. First, the transistor has an input conductance which is nonlinear and of appreciable magnitude in the normal operating frequency range of the device. Second, the device has internal coupling from output to input connections.

Ketchledge[1] has analyzed the production of amplitude distortion in feedback amplifiers containing one nonlinear element. Meyer[2] has investigated amplitude distortion in transistor amplifiers including a single-stage amplifier with feedback. It appears, however, that a detailed analysis of amplitude distortion in a transistor feedback amplifier of arbitrary complexity which includes the effects of nonlinearity in the transistor input circuit has not been reported.

In the present paper, attention is directed to the problem of calculating the distortion components of the collector current of a transistor (in the common-emitter connection) which is part of a feedback amplifier. Although the calculation process involves an approximation which is dependent on the input signal level, consideration is given to amplitude distortion arising both in the input circuit and in the forward transfer characteristic.

Equations for the distortion components associated with device models based on transistor input voltage and input current are derived using a theorem from feedback theory. The results obtained are considered for both ideal current and voltage signal sources as well as sources having arbitrary impedance characteristics. The effect on amplitude distortion as a result of the variation of the return ratio of the transistor with frequency is examined to provide guidelines for shaping this characteristic in the amplifier stabilization process. To illustrate the application of some of the equations derived, an example is presented in which the distortion components are computed for the output stage of a transistor feedback amplifier.

Determination of Distortion Products

METHOD OF ANALYSIS

The models shown in Figs. 1(A) and 1(B) will be used to account for the nonlinear behavior of a transistor when used in the common-emitter connection. It is known that the nonlinear behavior of a transistor amplifier stage is affected by the nature of the source which supplies its input signal, and this fact is relevant to the distortion analysis. A source having an output impedance which is low compared with the input impedance of the transistor approximates a constant-voltage source as seen by the transistor. The input voltage v_b is then a convenient variable for analysis, and the model of Fig. 1(A) is applicable. Similarly, when the amplifier stage is supplied from a source approximating a constant-current source, it is convenient to use the model of Fig. 1(B) and base the analysis of distortion on the variable i_b. It should be emphasized that the distortion coefficients a_{20}, a_{30}, b_{20}, and b_{30} do not have the same values in the two models shown in Fig. 1.

The central problem considered in this paper is the determination of the distortion components of the collector current when the transistor stage is part of a feedback amplifier whose input signal is sinusoidal. The results of the analysis will be presented both in terms of the input voltage (v_b) and input current (i_b) of the transistor. Providing information in the two forms allows a circuit designer to take advantage of the simplifications that result when the driving signal is approximated well by either a constant voltage or a constant current.

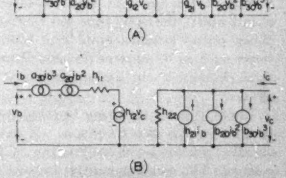


Fig. 1. Models for common-emitter transistor amplifier

A—Based on base-to-emitter voltage
B—Based on input current

Paper 61-97, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 7, 1960; made available for printing December 28, 1960.

J. H. MULLIGAN, JR., is with New York University, New York, N. Y.

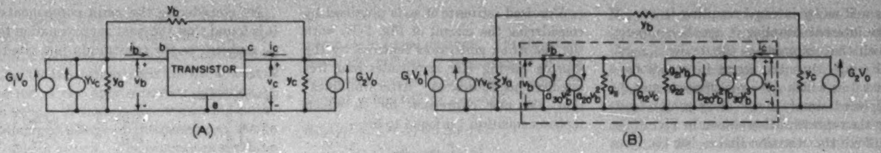


Fig. 2. Circuits for analysis

A—Transistor stage under consideration shown embedded in a nonreciprocal network
B—Equivalent circuit of the transistor combined with the network representing the remainder of the system

An essential part of the distortion calculation is the determination of the nonsinusoidal v_b (or i_b) by successive approximations. The actual input signal, that existing when the circuit is in its normal operating state, is determined by a transmission theorem from feedback theory.[3,4] The theorem states that the value of the input variable in the normal state is equal to the value of the same variable when the circuit is put in a reference state (that in which a certain nonreciprocal coupling is set equal to zero) divided by the return difference (F) for the particular nonreciprocal coupling.

The determination of distortion components rather than amplifier gain is of primary interest here. For this reason, it is convenient for analysis purposes to represent the portion of the amplifier external to the transistor by a nonreciprocal pi-network as indicated in Fig. 2(A). The complete amplifier, of which the stage being considered is only one part, is assumed to be supplied by a voltage source $V_0 \cos \omega t$. The effect of this source at base and collector ports is represented by the generators G_1V_0 and G_2V_0. It will be seen that the representation chosen simplifies the calculation of return ratio and return difference which are needed because of the use of the feedback theorem mentioned above. Naturally, if y_c is the load admittance for the complete amplifier and one wishes to compute the gain of the complete amplifier, an expression for gain may be obtained in terms of G_1, G_2, the elements of the pi-network, and the parameters of a model assumed for the transistor.

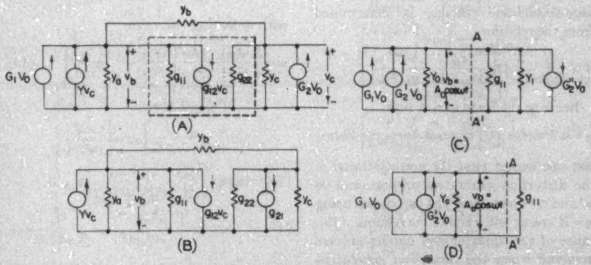


Fig. 3. Modified version of circuit in Fig. 2(A)

A—For computation of v_{b0}. The linear transistor model is contained within the dashed enclosure
B—For computation of $T(\omega)$
C—Equivalent circuit at base node for computation of $A_0(v_{b0})$ based on interpretation of equations 2 and 3:

$$Y_0 = y_a + \frac{y_b(g_{22}+y_c-Y)}{g_{22}+y_c+y_b}$$

$$Y_1 = \frac{g_{12}y_b}{g_{22}+y_c+y_b}$$

$$G_2' = \frac{G_2(Y+y_b)}{g_{22}+y_c+y_b}$$

$$G_2'' = \frac{G_2g_{12}}{g_{22}+y_c+y_b}$$

D—Approximate form of (C) which is useful in computation of v_b (although not necessarily i_b) when g_{12} is much less than Y

VOLTAGE-DEPENDENT MODEL

As has been indicated, in the analysis based on the model of Fig. 1(A), the voltage v_b will be computed by successive approximations. In the computation, the feedback theorem stated earlier will be used repeatedly, and the distortion generators shown in the model will be treated as independent sources in making each approximation.

First, v_b will be computed for the circuit using the voltage-dependent model, Fig. 2(B), in the linear condition (that in which $a_{20}=a_{30}=b_{20}=b_{30}=0$). To use the feedback theorem, let v_{b0} be the value of v_b when $g_{21}=0$. Then if

$$v_{b0} = A_0 \cos \omega t \tag{1}$$

v_b is given by*

$$v_b = A \cos \omega t \tag{2}$$

where

$$A = \frac{A_0}{F(\omega)} = \frac{A_0}{1+T(\omega)} \tag{3}$$

From Figs. 3(A) and 3(B), the quantities A_0 and $T(\omega)$ are found to be

$$A_0 = \frac{G_1V_0}{Y_{110}} + \frac{G_2V_0}{Y_{110}} \frac{(Y-g_{12}+y_b)}{(g_{22}+y_c+y_b)} \tag{4}$$

$$T(\omega) = \frac{g_{21}}{Y_{110}} \frac{(Y-g_{12}+y_b)}{(g_{22}+y_c+y_b)} \tag{5}$$

with

$$Y_{110} = g_{11} + y_a + \frac{y_b(g_{22}+y_c)}{g_{22}+y_c+y_b} + \frac{(Y-g_{12})y_b}{g_{22}+y_c+y_b} \tag{6}$$

In connection with the computation of A_0, it is of interest to note the circuit of Fig. 3(C) which is obtained by interpretation of equations 4 and 6. The elements to the left of the line A-A' represent the Thévenin equivalent source facing the base node of the transistor. It is seen in Fig. 3(A) that nonreciprocal coupling from collector to base nodes is provided by the circuit external to the device $[(Y+y_b)v_c]$

---

* To avoid an undesired complexity in the derivations, the phase angles associated with circuit transmissions will be neglected for the present. Thus, all quantities used are assumed to be real. Subsequently, the results obtained will be modified to take into account the complex nature of the impedances involved.

as well as by internal coupling ($g_{12}v_c$). If the internal coupling is small compared with that external to the device ($|g_{12}| \ll |Y+y_b|$), then the terms in equations 4 and 6 in which $(Y-g_{12})$ appears may be replaced simply by Y alone which leads to the representation shown in Fig. 3(D).

With the assertion that v_b has the form of equation 2, one may determine the contributions to the voltages and currents in the circuit of Fig. 2(B) produced by the base- and collector-circuit distortion generators raised to the second, third, and higher powers. Including the effect of these generators provides a more accurate estimate of v_b than equation 2 which considers only the effect of the input signal to the amplifier. Since frequencies other than ω result from the action of the distortion generators, however, equation 2 needs correction. To determine the effect of the distortion generators, their contribution to the base voltage under the condition $g_{21}=0$ can be determined and then the feedback theorem can be applied to find the value of v_b when the circuit is in its normal state. In this series of calculations, the distortion generators are taken as independent sources (with amplitudes dependent on the approximation to v_b when the circuit is in the normal state) acting in a linear system.

Consider now the effect produced on v_b because of the presence of the generators $a_{20}v_b{}^2$, $b_{20}v_b{}^2$, $a_{30}v_b{}^3$, and $b_{30}v_b{}^3$ in the base and collector circuits, respectively; see Fig. 2(B). Because of the identity

$$\cos^2 \omega t=\frac{1}{2}+\frac{1}{2}\cos 2\omega t$$

the assumption that v_b consists of the single term $A\cos\omega t$ results in the following current generators from the $v_b{}^2$ terms:

Base-circuit current generators:

$$\frac{1}{2}a_{20}A^2+\frac{1}{2}a_{20}A^2\cos 2\omega t$$

Collector-circuit current generators:

$$\frac{1}{2}b_{20}A^2+\frac{1}{2}b_{20}A^2\cos 2\omega t$$

Similarly, as a result of the equality

$$\cos^3 \omega t=\frac{3}{4}\cos\omega t+\frac{1}{4}\cos 3\omega t$$

the use of equation 2 results in third-order generators of the form:

Base-circuit current generators:

$$\frac{3}{4}a_{30}A^3\cos\omega t+\frac{1}{4}a_{30}A^3\cos 3\omega t$$

Collector-circuit current generators:

$$\frac{3}{4}b_{30}A^3\cos\omega t+\frac{1}{4}b_{30}A^3\cos 3\omega t$$

A revised estimate of v_{b0} is obtained by considering the circuit of Fig. 2(B) with the distortion generators replaced by the sources having the amplitudes and frequencies just indicated. After simplification, the new expression for v_{b0} which replaces equation 1 is found to be

$$\begin{aligned}v_{b0}=A_0\cos\omega t-\frac{1}{2}A^2\left[\frac{a_{20}}{Y_{110}}+\frac{b_{20}}{g_{21}}T(0)\right]-\\ \frac{1}{2}A^2\left[\frac{a_{20}}{Y_{110}}+\frac{b_{20}}{g_{21}}T(2\omega)\right]\cos 2\omega t-\\ \frac{3}{4}A^3\left[\frac{a_{30}}{Y_{110}}+\frac{b_{30}}{g_{21}}T(\omega)\right]\cos\omega t-\\ \frac{A^3}{4}\left[\frac{a_{30}}{Y_{110}}+\frac{b_{30}}{g_{21}}T(3\omega)\right]\cos 3\omega t\end{aligned} \tag{7}$$

Recall that this result is based on the first approximation to the base-to-emitter voltage when the circuit is in its normal condition ($v_b=A\cos\omega t$). Application of the feedback theorem to the result immediately above gives the second estimate of the voltage v_b when g_{21} has its normal value:

$$\begin{aligned}v_b=A\cos\omega t-\frac{A^2}{2}\left[\frac{a_{20}}{Y_{110}F(0)}+\frac{b_{20}}{g_{21}}\frac{T(0)}{F(0)}\right]-\\ \frac{A^2}{2}\left[\frac{a_{20}}{Y_{110}F(2\omega)}+\frac{b_{20}}{g_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\cos 2\omega t-\\ \frac{A^3}{4}\left[\frac{a_{30}}{Y_{110}F(3\omega)}+\frac{b_{30}}{g_{21}}\frac{T(3\omega)}{F(3\omega)}\right]\cos 3\omega t\end{aligned} \tag{8}$$

The value of A, the amplitude of the fundamental-frequency component of the base-to-emitter voltage, is determined from the relation:

$$A=\frac{A_0}{F(\omega)}-\frac{3}{4}A^3\left[\frac{a_{30}}{Y_{110}F(\omega)}+\frac{b_{30}}{g_{21}}\frac{T(\omega)}{F(\omega)}\right]$$

Because the form of v_b is now

$$v_b=a_0+a_1\cos\omega t+a_2\cos 2\omega t+a_3\cos 3\omega t$$

one can expect that the contributions of the distortion generators proportional to $v_b{}^2$ and $v_b{}^3$ which were just calculated using $v_b=A\cos\omega t$ will require correction. Because of the squaring and cubing process involved when the distortion generators are considered, v_b will be changed both because of the addition of new frequency components as well as the modification of the coefficients of the components contained in equations 7 and 8. In the present analysis of v_b, however, attention will be limited to the third harmonic and lower frequencies, and in addition, there will be included in the coefficients of the various terms only third or lower powers of A. Subject to these constraints, the only new terms *of interest* which are added to v_{b0} by the more accurate representation of v_{b2} are those arising from the current generators in base and collector circuits of value $a_{20}v_b{}^2$ and $b_{20}v_b{}^2$.

By considering the extra components, it is found that the final approximation to v_{b0} (subject to the constraints just cited) is

$$\begin{aligned}v_{b0}=A_0\cos\omega t-\\ A^3\left[\frac{3}{4}c_3(\omega)-c_2(\omega)\frac{c_2(0)}{F(0)}-\frac{1}{2}c_2(\omega)\frac{c_2(2\omega)}{F(2\omega)}\right]\times\\ \cos\omega t-\frac{1}{2}A^2c_2(0)-\frac{1}{2}A^2c_2(2\omega)\cos 2\omega t-\\ A^3\left[\frac{c_3(3\omega)}{4}-\frac{1}{2}c_2(3\omega)\frac{c_2(2\omega)}{F(2\omega)}\right]\cos 3\omega t\end{aligned} \tag{9}$$

where the coefficient $c_m(n\omega)$ is defined as follows (Y_{110} has been indicated as frequency-dependent in anticipation of the generalization of the results later in the development):

$$c_m(n\omega)=\frac{a_{m0}}{Y_{110}(n\omega)}+\frac{b_{m0}}{g_{21}}T(n\omega) \tag{10}$$

Use of the feedback transmission theorem on the terms in equation 9 gives a corrected version of v_b, the base-to-emitter voltage when g_{21} has its normal value:

$$\begin{aligned}v_b=A\cos\omega t-\frac{A^2}{2}\frac{c_2(0)}{F(0)}-\frac{A^2}{2}\frac{c_2(2\omega)}{F(2\omega)}\cos 2\omega t-\\ A^3\left[\frac{c_3(3\omega)}{4F(3\omega)}-\frac{1}{2}\frac{c_2(3\omega)}{F(3\omega)}\frac{c_2(2\omega)}{F(2\omega)}\right]\cos 3\omega t\end{aligned} \tag{11}$$

The quantity A is obtained by solving

$$A+mA^3=\frac{A_0}{F(\omega)} \tag{12}$$

with m having the value

$$m=\frac{3}{4}\frac{c_3(\omega)}{F(\omega)}-\frac{c_2(\omega)}{F(\omega)}\left[\frac{c_2(0)}{F(0)}+\frac{1}{2}\frac{c_2(2\omega)}{F(2\omega)}\right] \tag{13}$$

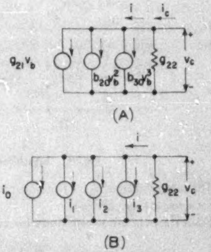


Fig. 4. Equivalent circuit at the collector node

A—Showing distortion generators added
B—With a d-c and three sinusoidal generators to account for fundamental and harmonic currents. Currents shown are those that exist when g_{21} has its normal value
$i_1=A_1\cos(\omega t+\theta_1)$
$i_2=A_2\cos(2\omega t+\theta_2)$
$i_3=A_3\cos(3\omega t+\theta_3)$

Consider now that, given A_0 and the linear transmission properties of the circuit, a solution for A has been obtained so that the base voltage v_b is known, at least in so far as the approximation equation 11 is valid. Attention is now directed to the combined effects of the generators in the collector circuit. The generators are those which have been shown in the circuits considered earlier; they are represented again in the equivalent collector circuit of Fig. 4(A). The total current flowing into the collector terminal is i_c; it is the sum of the current in g_{22} plus that taken by the three current generators (i). By inspection,

$$i = g_{21}v_b + b_{20}v_b^2 + b_{30}v_b^3 \tag{14}$$

Limiting attention to frequency components equal to or less than 3ω and remembering that only powers of A up to and including the third are to be included, one can write

$$i = i_0 + i_1 + i_2 + i_3 \tag{15}$$

to represent the sum of the d-c component (i_0) and frequency components of ω, 2ω, and 3ω. The individual current components are determined by using equation 11 in equation 14 and simplifying using equations 10, 12, and 13. The final expressions for the components are

$$i_0 = \frac{A^2}{2F(0)}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(0)}\right] \tag{16}$$

$$i_1 = \left[g_{21}\frac{A_0}{F(\omega)} + \frac{A^3}{F(\omega)}\frac{3}{4}\left(b_{30} - \frac{g_{21}a_{30}}{Y_{110}(\omega)}\right) - \frac{A^3}{F(\omega)}\left(b_{20} - \frac{g_{21}a_{20}}{Y_{110}(\omega)}\right)\left\{\frac{b_{20}}{g_{21}}\left[\frac{T(0)}{F(0)} + \frac{1}{2}\frac{T(2\omega)}{F(2\omega)}\right] + a_{20}\left[\frac{1}{Y_{110}(0)F(0)} + \frac{1}{2Y_{110}(2\omega)F(2\omega)}\right]\right\}\right]\cos\omega t \tag{17}$$

$$i_2 = \frac{A^2}{2F(2\omega)}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(2\omega)}\right]\cos 2\omega t \tag{18}$$

$$i_3 = \frac{A^3}{F(3\omega)}\left\{\frac{1}{4}\left[b_{30} - \frac{g_{21}a_{30}}{Y_{110}(3\omega)}\right] - \frac{1}{2}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(3\omega)}\right]\left[\frac{a_{20}}{Y_{110}(2\omega)F(2\omega)} + \frac{b_{20}}{g_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\right\}\cos 3\omega t \tag{19}$$

Equations 16 through 19 give the strengths of the current generators in the collector circuit illustrated in Fig. 4(B). The results have been obtained without regard to phase shifts in the transmission paths, and it is desired now to present a more general result. Let the voltage at the base in the condition $g_{21}=0$ be identified as

$$v_{b0} = A_0 \cos\omega t \tag{1}$$

where A_0 is a real number. On the other hand, the voltage of frequency ω at the same point in the system when g_{21} has its normal value will be denoted

$$R_e[A_f e^{j\theta} e^{j\omega t}] = A_f \cos(\omega t + \theta)$$

The values of A_f and θ are determined from

$$A_f e^{j\theta} + \left\{\frac{3}{4}\left[\frac{a_{30}}{Y_{110}(\omega)F(\omega)} + \frac{b_{30}}{g_{21}}\frac{T(\omega)}{F(\omega)}\right] - \left[\frac{a_{20}}{Y_{110}(\omega)F(\omega)} + \frac{b_{20}}{g_{21}}\frac{T(\omega)}{F(\omega)}\right] \times \left[\frac{a_{20}}{Y_{110}(0)F(0)} + \frac{b_{20}}{g_{21}}\frac{T(0)}{F(0)}\right] + \frac{1}{2}\left[\frac{a_{20}}{Y_{110}(2\omega)F(2\omega)} + \frac{1}{2}\frac{b_{20}}{g_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\right\} \times A_f^3 e^{j\theta} = \frac{A_0}{F(\omega)} \tag{20}$$

The quantities Y_{110}, F, and T are complex; the remaining quantities are real. When A_f and θ have been determined, the generators in Fig. 4(B) may be identified as follows:

$$i_0 = \frac{A_f^2}{2F(0)}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(0)}\right] \tag{21}$$

$$i_1 = R_e\left\{\left[\frac{g_{21}A_0}{F(\omega)} + \frac{A_f^3 e^{j\theta}}{F(\omega)}\frac{3}{4}\left[b_{30} - \frac{g_{21}a_{30}}{Y_{110}(\omega)}\right] - \frac{A_f^3 e^{j\theta}}{F(\omega)}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(\omega)}\right]\left\{\frac{b_{20}}{g_{21}}\left[\frac{T(0)}{F(0)} + \frac{1}{2}\frac{T(2\omega)}{F(2\omega)}\right] + a_{20}\left[\frac{1}{Y_{110}(0)F(0)} + \frac{1}{2Y_{110}(2\omega)F(2\omega)}\right]\right\}\right]e^{j\omega t}\right\} \tag{22}$$

$$i_2 = R_e\left\{\frac{1}{2F(2\omega)}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(2\omega)}\right]A_f^2 e^{j2\theta} e^{j2\omega t}\right\} \tag{23}$$

$$i_3 = R_e\left[\left\{\frac{1}{4}\left[b_{30} - \frac{g_{21}a_{30}}{Y_{110}(3\omega)}\right] - \frac{1}{2}\left[b_{20} - \frac{g_{21}a_{20}}{Y_{110}(3\omega)}\right]\left[\frac{a_{20}}{Y_{110}(2\omega)F(2\omega)} + \frac{b_{20}}{g_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\right\}\frac{A_f^3 e^{j3\theta}}{F(3\omega)}e^{j3\omega t}\right] \tag{24}$$

It is to be noted that the return differences and return ratios that appear in the equations above are those computed for g_{21} for zero reference. Furthermore, it is of interest to observe that, in general, there is an additional contribution to v_c as a result of transmission from the base-circuit-distortion generators to the collector circuit through paths that do not depend on g_{21}. In most practical situations, however, the transmission of this nature is quite small by comparison with that involving g_{21} which is calculated above.

Finally, it is to be emphasized that equation 20 gives only an approximation to the value of A_f. The combination of signal level (A_0/F), circuit transmission, and distortion coefficients may cause the term proportional to A_f^3 to be comparable in amplitude with A_f itself. If this occurs, equation 20 may not describe the circuit operation accurately, and one must include the effect of distortion processes yielding $\cos\omega t$ terms with amplitudes proportional to A_f^5 (or higher powers of A_f). If the distortion products are large enough to cause this effect, it becomes apparent when one tries to solve the equation for A_f.

CURRENT-DEPENDENT MODEL

The transistor will be represented by the model of Fig. 1(B), and as before, it will be taken as embedded in a network represented by a nonreciprocal pi-structure. The composite circuit which is to be analyzed is shown in Fig. 5(A). Corresponding to the earlier transistor analysis, the objective of the analysis is to determine the current i_c when the circuit is in its normal state. The procedure will follow that used previously, i.e., an expression for i_b will be obtained from which the contributions of the individual collector-circuit generators can be determined. Computation of i_b will be made applying the feedback transmission theorem used above. This implies that i_{b0}, the amplitude of i_b in the condition $h_{21}=0$, will be evaluated and divided by the return difference for h_{21} in order to find the amplitude of i_b when h_{21} has its normal value.

A circuit from which the return difference for h_{21} may be found is shown in Fig. 5(B). The return ratio, from which F is found by adding unity, is the negative of the (complex) amplitude of i_b in this circuit. It is found by direct calculation that

$$T(\omega) = \frac{h_{21}}{h_{11}}\frac{[Y + y_b - h_{12}(y_a + y_b)]}{Y_{110}(h_{22} + y_b + y_c)} \tag{25}$$

where Y_{110} is the total driving point admittance at the base node when $h_{21}=0$, viz.,

$$Y_{110} = y_a - \frac{y_b(h_{22} + y_c + Y)}{h_{22} + y_c + y_b} + \frac{1}{h_{11}}\left[1 - \frac{h_{12}y_b}{h_{22} + y_b + y_c}\right] \tag{26}$$

Let the base current when $h_{21}=0$ be denoted as

$$i_{b0} = A_0 \cos\omega t \tag{27}$$

Equation 27 is the value which i_b has in the circuit of Fig. 5(C). It is found from the circuit shown in that figure that the value of A_0 is given by

$$A_0=\frac{G_1V_0}{h_{11}Y_{11}}\left[1-\frac{h_{12}y_b}{h_{22}+y_b+y_c}\right]+\frac{1}{h_{21}}G_2V_0T(\omega) \qquad (28)$$

The value of i_b when the circuit is in the normal state is then

$$i_b=A\cos\omega t \qquad (29)$$

where

$$A=A_0/F(\omega) \qquad (30)$$

From the previous analysis, it is realized that equation 29 represents only a first approximation to the current i_b. To obtain a more accurate result, the effects of the base- and collector-circuit distortion generators must be included. As before, the technique for including these effects will be to assume that i_b is given by equation 29 for the purpose of determining the strengths of the distortion generators in base and collector circuits. To find the contributions made to i_b by these sources, it is necessary to determine the components of i_b they yield for $h_{21}=0$ and then apply the feedback theorem to ascertain the net effect on the system in the normal state.

If the effects of generators $a_{20}i_b^2$, $a_{30}i_b^3$, $b_{20}i_b^2$, and $b_{20}i_b^3$ are taken into account, the value of i_{b0} is modified to

$$\begin{aligned}i_{b0}=A_0\cos\omega t-\frac{1}{2}A^2\left[\frac{a_{20}}{Z(0)}+\frac{b_{20}}{h_{21}}T(0)\right]-\\ \frac{1}{2}A^2\left[\frac{a_{20}}{Z(2\omega)}+\frac{b_{20}}{h_{21}}T(2\omega)\right]\cos 2\omega t-\\ \frac{3A^3}{4}\left[\frac{a_{30}}{Z(\omega)}+\frac{b_{30}}{h_{21}}T(\omega)\right]\cos\omega t-\\ \frac{A^3}{4}\left[\frac{a_{30}}{Z(3\omega)}+\frac{b_{30}}{h_{21}}T(3\omega)\right]\cos 3\omega t\end{aligned} \qquad (31)$$

where

$$Z(\omega)=h_{11}+Z_0\left[1-\frac{h_{12}y_b}{h_{22}+y_c+y_b}\right] \qquad (32)$$

By application again of the feedback theorem, equation 31 may be utilized to obtain the form of i_b when the circuit is in its normal state. Thus, with h_{21} having its normal value, there is obtained:

$$\begin{aligned}i_b=A\cos\omega t-\frac{A^2}{2}\left[\frac{a_{20}}{Z(0)F(0)}+\frac{b_{20}}{h_{21}}\frac{T(0)}{F(0)}\right]-\\ \frac{A^2}{2}\left[\frac{a_{20}}{Z(2\omega)F(2\omega)}+\frac{b_{20}}{h_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\cos 2\omega t-\\ \frac{A^3}{4}\left[\frac{a_{30}}{Z(3\omega)F(3\omega)}+\frac{b_{30}}{h_{21}}\frac{T(3\omega)}{F(3\omega)}\right]\cos 3\omega t\end{aligned} \qquad (33)$$

The amplitude A of the component of frequency ω is found by solving the equation

$$A+mA^3=\frac{A_0}{F(\omega)}$$

where $F(\omega)$ is the return difference for h_{21} (zero reference) and m is

$$m=\frac{3}{4}\left[\frac{a_{30}}{Z(\omega)F(\omega)}+\frac{b_{30}}{h_{21}}\frac{T(\omega)}{F(\omega)}\right]$$

From the earlier work with the model of Fig 1(A), it is realized that equation 33 must be modified if one wishes to include amplitude terms proportional to A^3 (but not higher) in the expression containing frequency components up to 3ω. Additional terms result from the combination of the d-c and $\cos\omega t$ as well as the $\cos\omega t$ and $\cos 2\omega t$ factors in the operations indicated by $a_{20}i_b^2$ and $b_{20}i_b^2$. Ultimately, also, one must include the complex nature of the various transmissions involved to obtain the proper phase relationships of the various currents. Full details of the process involved in modifying the equations to provide these refinements were given in connection with the prior analysis of the voltage-controlled transistor model. By noting the similarity of the present analysis to the preceding formulation, it is possible to proceed directly to the final results.

Fig. 6(A) shows the equivalent collector circuit for the device model under consideration with the distortion generators $b_{20}i_b^2$ and $b_{30}i_b^3$ included. The counterpart of equation 14 may be written as

$$i=h_{21}i_b+b_{20}i_b^2+b_{30}i_b^3 \qquad (34)$$

Now, if as in the preceding analysis, the components of i are limited to direct current and frequencies ω, 2ω, and 3ω, then equation 15 may be repeated, applying to Fig. 6(B) in the present instance, i.e.,

$$i=i_0+i_1+i_2+i_3 \qquad (15)$$

The signal at the base in the condition $h_{21}=0$ is taken as the reference for the phasor calculations. Assume

$$i_{b0}=A_0\cos\omega t$$

where A_0 is a real number. The component of frequency ω in i_b when h_{21} has its normal value is written as

$$R_e[A_f\epsilon^{j\theta}\epsilon^{j\omega t}]=A_f\cos(\omega t+\theta)$$

The quantities A_f and θ must be obtained, in general, as solutions of the equation

$$\begin{aligned}A_f\epsilon^{j\theta}+\Bigg\{\frac{3}{4}\left[\frac{a_{30}}{Z(\omega)F(\omega)}+\frac{b_{30}}{h_{21}}\frac{T(\omega)}{F(\omega)}\right]-\\ \left[\frac{a_{20}}{Z(\omega)F(\omega)}+\frac{b_{20}}{h_{21}}\frac{T(\omega)}{F(\omega)}\right]\times\\ \left[\frac{a_{20}}{Z(0)F(0)}+\frac{b_{20}}{h_{21}}\frac{T(0)}{F(0)}+\frac{1}{2}\frac{a_{20}}{Z(2\omega)F(2\omega)}+\right.\\ \left.\frac{1}{2}\frac{b_{20}}{h_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\Bigg\}A_f\epsilon^{3j\theta}=\frac{A_0}{F(\omega)}\end{aligned} \qquad (35)$$

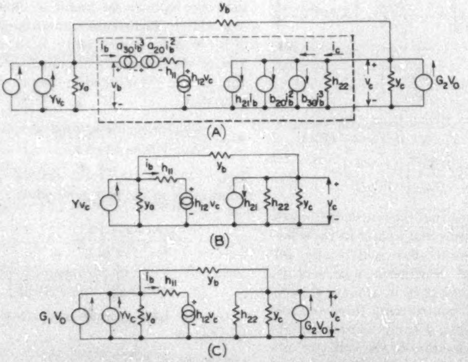


Fig. 5. Transistor stage represented by model of Fig. 1(B)

A—Shown embedded in a nonreciprocal network
B—Simplified form used to calculate $T(\omega)$ for h_{21}
C—Simplified form used to determine i_b under the condition $h_{21}=0$

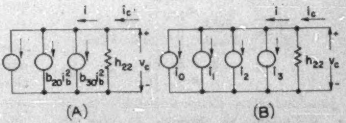


Fig. 6. Equivalent circuit at the collector node

A—Showing distortion generators added
B—With a d-c and three sinusoidal generators to account for fundamental and harmonic currents. Currents shown are those that exist when h_{21} has its normal value
$i_1=A_1\cos(\omega t+\theta_1)$
$i_2=A_2\cos(2\omega t+\theta_2)$
$i_3=A_3\cos(3\omega t+\theta_3)$

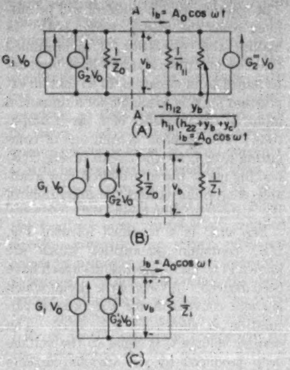


Fig. 7. Equivalent circuits for the base node based on interpretation of equation 30

A—Without approximation, where

$$\frac{1}{Z_0} = y_e + \frac{y_b(h_{22}+y_e-Y)}{h_{22}+y_e+y_b}$$

$$G_2' = \frac{G_2(Y+y_b)}{h_{22}+y_b+y_e}$$

$$G_2''' = \frac{h_{12}G_2}{h_{11}(h_{22}+y_e+y_b)}$$

B—An approximation which is useful when $|Y+y_b| \gg |h_{12}/h_{11}|$

$$\frac{1}{Z_1} = \frac{1}{h_{11}}\left[1 - \frac{h_{12}y_b}{h_{22}+y_b+y_e}\right]$$

C—An approximation which is useful when $|Y+y_b| \gg |h_{22}/h_{11}|$ and $Z_0 \gg Z_1$. The value of A_0 is the magnitude of the quantity

$$G_1V_0 + \frac{G_2V_0(Y+y_b)}{h_{22}+y_b+y_e}$$

The quantities $F(n\omega)$, $T(n\omega)$ are the return difference and return ratio for h_{21}. In general, they and the quantity $Z(n\omega)$ are complex. Once A_f and θ have been determined, however, the individual components of i shown in Fig. 6(B) may be evaluated. (The comments regarding the approximation to A_f made in connection with equation 20 apply equally well to equation 35.) Expressions for these are

$$i_0 = \frac{A_f^2}{2F(0)}\left[b_{20} - \frac{h_{21}a_{20}}{Z(0)}\right] \tag{36}$$

$$i_1 = R_e\left\{\left[\frac{h_{21}A_0}{F(\omega)} + \frac{A_f^3 e^{j\theta}}{F(\omega)}\frac{3}{4}\left[b_{30} - \frac{h_{21}a_{30}}{Z(\omega)}\right] - \frac{A_f^3 e^{j\theta}}{F(\omega)}\left[b_{20} - \frac{h_{21}a_{20}}{Z(\omega)}\right] \times \left\{\frac{b_{20}}{h_{21}}\left[\frac{T(0)}{F(0)} + \frac{1}{2}\frac{T(2\omega)}{F(2\omega)}\right] + a_{20}\left[\frac{1}{Z(0)F(0)} + \frac{1}{2}\frac{1}{Z(2\omega)F(2\omega)}\right]\right\}\right] e^{j\omega t}\right\} \tag{37}$$

$$i_2 = R_e\left\{\frac{1}{2F(2\omega)}\left[b_{20} - \frac{h_{21}a_{20}}{Z(2\omega)}\right] A_f^2 e^{j2\theta} e^{j2\omega t}\right\} \tag{38}$$

$$i_3 = R_e\left[\left\{\frac{1}{4}\left[b_{30} - \frac{h_{21}a_{30}}{Z(3\omega)}\right] - \frac{1}{2}\left[b_{20} - \frac{h_{21}a_{20}}{Z(3\omega)}\right] \left[\frac{a_{20}}{Z(2\omega)F(2\omega)} + \frac{b_{20}}{h_{21}}\frac{T(2\omega)}{F(2\omega)}\right]\right\} \frac{A_f^3 e^{j3\theta}}{F(3\omega)e^{j3\omega t}}\right] \tag{39}$$

The family of equations, 36 to 39, provides the values of the collector-circuit current generators of Fig. 6(B). This set of equations takes into account the effect of internal feedback in the transistor on the collector currents, but the contribution of the base-circuit generators to v_c, and hence to i_c, by transmission through y_b which is independent of h_{21} has not been included. Similar statements apply to the corresponding set of equations, 21 to 24, for the voltage-dependent model. It should be emphasized that F and T which appear in the equations immediately above are for h_{21} (for zero reference) whereas the same symbols which appear in the set, equations 21 to 24, are for g_{21} for the same reference.

APPLICABILITY OF RESULTS

Now that two sets of equations are available corresponding to the choice of base voltage or base current as the independent variable, it is helpful to consider the applicability of the two solutions under certain limiting conditions of transistor operation. Equation 28, which is used for the determination of the amplitude of the base current when $h_{21}=0$, may be interpreted in terms of the equivalent circuit given in Fig. 7(A). The transistor input circuit is shown to the right of the dashed line A-A'. The effect at the base node of the source G_2V_0 is represented by two current generators, one internal to the transistor and one external to it. In so far as the nodal voltage v_b is concerned, if the internal coupling proportional to h_{12}/h_{11} is small compared with the external feedback through $Y+y_b$, that is, if $|Y+y_b| \gg |h_{12}/h_{11}|$, then the circuit of Fig. 7(B) is a good approximation to the exact representation in Fig. 7(A). Finally, if in addition, the transistor is embedded in a network such that the equivalent impedance Z_0 facing it is much greater than Z_1, the input impedance to the amplifier stage with $h_{21}=0$, then the circuit of Fig. 7(C) is applicable. The determination of A_0 under these conditions is clearly a simple matter since all the current from the equivalent-current source is delivered to the base. With A_0 known, one may proceed to the calculation of collector-current components utilizing equations 36–39.

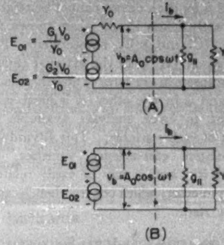

Y_0
i_b
$E_{01} = \frac{G_1 V_0}{Y_0}$
$E_{02} = \frac{G_2' V_0}{Y_0}$
$v_b = A_0 \cos \omega t$
g_{11}
Y_1
(A)
E_{01}
E_{02}
(B)

Fig. 8. Circuit equivalent to that of Fig. 3(C)

A—Approximate form valid when $|Y+y_b| \gg |g_{12}|$

B—An approximate form which is useful when $1/Y_0$ is much less than the impedance seen to the right of the dashed section (approximately, the condition is $Y_0 \gg g_{11}$). The amplitude of A_0 is equal to the magnitude of

$$\frac{G_1V_0}{Y_0} + \frac{G_2'V_0}{Y_0}$$

Suppose that the impedance facing the transistor is low compared with the transistor input impedance when the circuit is in the condition $h_{21}=0$. Then it is convenient to deal with v_b rather than i_b. The convenience arises because the base-to-emitter voltage is approximately equal to the open-circuit voltage of the source. Fig. 3(C) corresponds to Fig. 7(A) except that the former is based on v_b as the independent variable for the distortion analysis. The circuit external to the transistor (to the left of the dashed line) may be written in the form shown in Fig. 8(A). If $|Y+y_b| \gg |g_{12}|$, then the internal generator depending on G_2V_0 may be neglected giving the representation shown to the right of the dashed line in Fig. 8(A). Subject to the further condition that $1/Y_0$ [see Fig. 3(C)] is much less than the input impedance of the transistor in the state $g_{21}=0$, conditions at the base node are represented approximately by the circuit of Fig. 8(B). It is seen from the last representation that A_0 is approximately equal to the amplitude of the open-circuit voltage of the source connected to the base when $Y_0 \gg g_{11}$. With A_0 established, the distortion current components in the collector circuit can be computed using equations 21–24.

It is of interest to observe from Figs. 7(C) and 8(B) that v_b and i_b are not simply related unless $g_{12}=h_{12}=0$ (condition of no internal feedback). In particular, even in the limiting conditions of operation depicted by the two figures,

v_b and i_b for each device cannot be related accurately unless h_{12} and g_{12} are known, for Figs. 7(C) and 8(B), respectively. For this reason, i.e., since the external circuit determines v_b in one form of limiting condition and i_b in the other, it is most convenient to have analyses available in terms of each variable. This advantage of relative convenience vanishes, of course, if v_b and i_b may be easily related to one another, which occurs when g_{12} and h_{12} are zero.

SIMPLIFICATION WITH IDEAL SOURCES

The equations for the distortion components take on a particularly simple form if the driving signal for the transistor stage comes from either a constant-current source or a constant-voltage source. Although neither of these ideal conditions may be attained physically, of course, practical circuit conditions sometimes may be approximated well by one of the two extremes. The assumption of a constant-voltage source implies that Y_0 of Figs. 3(C) and 8(A) becomes infinite with the consequence that the quantity $Y_{110}(\omega)$ appearing in the expressions derived in terms of v_b also becomes infinite. The assumption of a constant-current source similarly implies that Z_0 of Fig. 7(A) becomes infinite. As a result, the impedance $Z(\omega)$ which is contained in the equations which were derived for the current-controlled transistor model also increases without limit. Inasmuch as the families of equations which were derived for the two transistor models are so similar in form, it is found that under the assumed limiting conditions of a constant-voltage source or constant-current source, the equations for both situations reduce to the form:

$$i_0=\frac{1}{2}\,b_{20}\,\frac{A_f^2}{F(0)} \tag{40}$$

$$i_1=R_e\left[\left\{\frac{xA_0}{F(\omega)}+\frac{1}{F(\omega)}\left[\frac{3}{4}b_{30}-\frac{b_{20}^2}{x}\,\frac{T(0)}{F(0)}-\frac{1}{2}\,\frac{b_{20}^2}{x}\,\frac{T(2\omega)}{F(2\omega)}\right]A_f^3\epsilon^{j\theta}\right\}\epsilon^{j\omega t}\right] \tag{41}$$

$$i_2=R_e\left[\frac{1}{2}\,b_{20}\,\frac{A_f^2}{F(2\omega)}\,\epsilon^{j2\theta}\epsilon^{2\omega t}\right] \tag{42}$$

$$i_3=R_e\left\{\left[\frac{1}{4}b_{30}-\frac{1}{2}\,\frac{b_{20}^2}{x}\,\frac{T(2\omega)}{F(2\omega)}\right]\frac{A_f^3\epsilon^{j3\theta}}{F(3\omega)}\,\epsilon^{j3\omega t}\right\} \tag{43}$$

In these equations A_f and θ are determined from an equation of the form

$$A_f\epsilon^{j\theta}+\left\{\frac{3}{4}\,\frac{b_{30}}{x}-\frac{b_{20}^2}{x^2}\left[\frac{T(0)}{F(0)}+\frac{1}{2}\,\frac{T(2\omega)}{F(2\omega)}\right]\right\}\times\frac{T(\omega)}{F(\omega)}\,A_f^3\epsilon^{j\theta}=\frac{A_0}{F(\omega)} \tag{44}$$

When this family of equations is applied to the limiting constant-current source portrayed in Fig. 7(C), the variable x is h_{21}, the return differences and return ratios F and T are computed for h_{21}, and the coefficients b_{20} and b_{30} are those indicated in Fig. 1(B). On the other hand, when the set of relations is used as an approximation to the constant-voltage source depicted in Fig. 8 (B), the variable x is g_{21}, the return differences and return ratios F and T are computed for g_{21}, and the coefficients b_{20} and b_{30} are the ones illustrated in Fig. 1(A).

DETERMINATION OF OUTPUT CURRENT AND VOLTAGE

A problem of practical interest is the calculation of signal voltage at the collector (v_c) or the signal current to the collector (i_c) in terms of the components i_0, i_1, i_2, and i_3 which have been considered. The solution to this problem will be outlined using the model of Fig. 1(A). In making the calculation, it is presumed that the quantities A_f and θ have been determined from equation 20 so that an expression for v_b is available for a sinusoidal input to the amplifier of a given amplitude. The problem is then one of computing v_c or i_c for the circuit of Fig. 2(B) making use of the information now available concerning v_b and the amplitude of the distortion generators.

Fig. 9(A) is a modified form of Fig. 2(B) containing quantities which are known in terms of expressions which have been derived. A form of the circuit which is more convenient for the present calculation is shown in Fig. 9(B). The desired voltage v_c is equal to the voltage drop produced by the various currents indicated flowing through the admittance $g_{22}+y_b+y_c$. The collector current i_c is the sum of $g_{22}v_c$ and i.

In many practical transistor feedback amplifiers $G_2V_0\doteq 0$ and $|g_{21}|\gg|y_b|$, so that Fig. 9(B) reduces to the approximate form shown in Fig. 9(C). For this configuration, v_c and i_c may be determined by inspection. For example, one finds by using equation 23 that the second-harmonic current in y_c (reference direction downward) is

$$i_{2c}=-R_e\left\{\frac{y_c}{g_{22}+y_c+y_b}\left[b_{20}-\frac{g_{21}a_{20}}{Y_{110}(2\omega)}\right]\times\frac{A_f^2}{2F(2\omega)}\,\epsilon^{j2\theta}\epsilon^{j2\omega t}\right\} \tag{45}$$

where use has been made of equation 23 for i_2.

Interpretation of Results

DISTORTION WITH IDEAL SOURCES

In certain amplifier circuits the relative values of transistor input impedance and impedance of the source driving the amplifying stage are such that a good approximation to input terminal current or voltage is obtained by assuming either a constant-current or constant-voltage drive. Equations 40 to 44 are quite useful for distortion calculations in these limiting conditions; they also provide a convenient starting point for a discussion of the effect of feedback on amplitude distortion.

The analysis in this paper is limited to the inclusion of terms containing the base-signal amplitude raised to the third or lower powers. Because of this restriction, the expressions for the d-c component i_0 and the second harmonic i_2 have the same form, both being proportional to $b_{20}A_f^2$ and inversely proportional to the

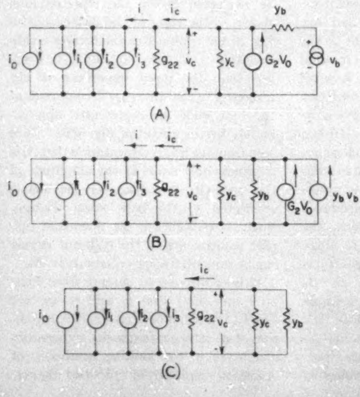


Fig. 9. Combination of Figs. 2(B) and 4(B)

A—For the computation of v_c when i and v_b are known
B—Equivalent form of (A)
C—Approximation to (B) which is valid when $|g_{21}|\gg|y_b|$ and $G_2V_0\doteq 0$

value of F at the frequency of the component. Thus, the value of $F(0)$, the return difference at direct current, is of great importance in establishing the amplitude of i_0. If the feedback path has a low-pass transmission characteristic so that appreciable loop gain is maintained through the low-frequency spectrum down to direct currents, i_0 will be considerably smaller than its nonfeedback counterpart would be with the same base-circuit signal level. When the feedback loop contains a-c coupled circuits, however, the return ratio becomes vanishingly small as the frequency is reduced. When this occurs, the component i_0 for the transistor in the feedback amplifier experiences no reduction relative to the analogous component for the transistor in the circuit without feedback.

The amplitude of i_2 is similarly affected by the value of $F(2\omega)$. If ω is in the useful operating frequency range of the amplifier, $|T(\omega)|$ will usually be much greater than unity and constant. Therefore, no significant change will occur in the amplitude of i_2 as ω is increased, although its phase angle, directly dependent on that of $F(\omega)$, will undergo appreciable variation with frequency. As the frequency ω is increased to the point where $|T(2\omega)|$ approaches unity, the value of $1/F(2\omega)$ becomes quite sensitive to the angle of $T(2\omega)$. The amplitude of i_2 can then be increased considerably with respect to its value in the operating frequency range; indeed, in this cutoff region, the amplitude may be increased because of the variation of $F(2\omega)$ to values greater than obtained in the absence of feedback. The actual enhancement depends on the values assumed by the return ratio in the region where its magnitude is near unity.

Similar possibilities for enhancement are seen in the expression for i_3. As $T(3\omega)$ approaches the region where its magnitude is near unity, an increase in amplitude of i_3 is to be expected which is dependent on the angle of T. If one considers that frequency is raised from a value where 2ω and 3ω are both in the region where $|T(\omega)|$ is essentially constant, then it would be expected that the relative amplitudes of i_2 and i_3 would change as 3ω, and subsequently 2ω, become located in the region where $|T|$ passes through unity. It is seen from equation 43, however, that the variation of i_3 is somewhat more involved than a mere dependence on $A_f^3/F(3\omega)$. The production of third harmonic as a result of the interaction of the distortion generated cos $2\omega t$ with the fundamental causes the presence of the term

$$-\frac{1}{2}\frac{b_{20}}{x}\frac{T(2\omega)}{F(2\omega)}$$

It may be shown that if $|T(\omega)| \geq 6$, the magnitude of the ratio T/F is limited to the range 0.86 to 1.20 whereas its angle must lie between -10 and $+10$ degrees. Unless $|T(2\omega)|$ is near unity, therefore, one may consider in the interpretation of the equations that $T(2\omega)/F(2\omega)$ is approximately unity without committing a serious error. Thus, for frequencies such that $|T(2\omega)| >> 1$, the factor multiplying $A_f^3/F(3\omega)$ is approximately equal to

$$\frac{b_{30}}{4}-\frac{1}{2}\frac{b_{20}^2}{x}$$

where x is either h_{21} or g_{21}. If b_{30} is positive, the feedback associated with the second-order distortion products results in a smaller third-harmonic coefficient than in the corresponding nonfeedback operation. The actual third-harmonic *component* is, of course, still further affected by the factor $1/F(3\omega)$ which, in most practical situations, will cause considerable reduction of the third-harmonic distortion component as compared with its value in the absence of feedback.

The desired fundamental-frequency component of collector current is $xA_0/F(\omega)$. Unfortunately, three other components appear in equation 41 for i_1 because of the distortion process; all are proportional to $A_f^3/F(\omega)$. Two of the terms occur because of second-order distortion products combining with fundamental; they contain the ratio T/F at direct current and 2ω. Since these two terms are negative with respect to $xA_0/F(\omega)$, they are responsible for *compression* of the fundamental. The third term is proportional to b_{30}. If this coefficient is positive, *expansion* will be present in the nonfeedback version of the amplifier. In the presence of feedback, however, the compression caused by the second-order products may permit partial or complete cancellation of this otherwise detrimental effect. On the other hand, if b_{30} is negative, the compression *coefficient* with feedback is worse than in the nonfeedback situation because of the additional contribution from the second-order products. Since A_f is the magnitude of the base-circuit signal, and since the net compression or expansion is proportional to $A_f^3/F(\omega)$, the net nonlinearity with feedback where $|T(\omega)| >> 1$ will always be considerably less than without. It is noted particularly that one of the second-order terms is

$$\frac{b_{20}^2}{x}\frac{T(0)}{F(0)}$$

If there is local feedback at direct current, which might occur because of bias circuits, for example, $T(0)$ will not be zero and some compression will result. This can occur even though the amplifier might normally be viewed as a nonfeedback amplifier because of the absence of a main feedback loop and even though the bypassing of bias networks over the normal operating frequency range of the amplifier is adequate.

DISTORTION WITH ARBITRARY SOURCE IMPEDANCE

When circuit conditions exist which make the assumption of an idealized drive condition untenable, the effect of the base-circuit distortion generators will be observed in the expressions for the components of collector current. The behavior to be expected can be predicted by a consideration of the sets of equations 21 to 44 and 36 to 39.

The effect of base-circuit distortion on i_0 and i_2 is relatively easy to assess inasmuch as the coefficients by which A_f^2 must be multiplied are affected in a simple fashion. A knowledge of a_{20} for the device model selected and either $Z(\omega)$ or $Y_{110}(\omega)$ permits an accurate estimate to be made rather quickly. If, for example, a_{20} and b_{20} have the same sign and $Y_{110}(\omega)$ or $Z(\omega)$ are essentially resistive, one can state that the total distortion components i_0 and i_2 tend to be reduced rather than increased by permitting base-circuit harmonics to affect collector current. The possibility of coefficient reduction by partial or complete cancellation of b_{20} by the terms

$$\frac{g_{21}a_{20}}{Y_{110}}$$

or

$$\frac{h_{21}a_{20}}{Z}$$

is quite apparent. It is noted in connection with i_2 that the modification of b_{20} will change with frequency if $Y_{110}(2\omega)$ and $Z(2\omega)$ are frequency-dependent.

The pattern of change is not so obvious, however, when one considers the structure of i_1 and i_3. To be sure, b_{20} and b_{30} are each modified by the subtraction of a term containing a_{20} or a_{30}. Were this the only effect, its consequences could be forecast to some degree as was done in connection with i_0 and i_2. In i_3, however, one must consider, in addition, the importance of the factor

$$\frac{a_{20}}{Y_{110}(2\omega)F(2\omega)}$$

or

$$\frac{a_{20}}{Z(2\omega)F(2\omega)}$$

which is added to

$$\frac{b_{20}}{g_{21}}\frac{T(2\omega)}{F(2\omega)}$$

or

$$\frac{b_{20}}{h_{21}}\frac{T(2\omega)}{F(2\omega)}$$

For a complete discussion, one would need to have available knowledge concerning Y_{110} and Z. Assuming that these remain essentially resistive, it appears that the importance of the added terms containing a_{20} will increase as the signal frequency moves out of the normal operating range where $T(\omega)$ is assumed to have a constant magnitude and into the cutoff region where $T(\omega)$, and hence $F(\omega)$, are expected to decrease. In particular, if Y_{110} and Z have not increased significantly by the time $|T(2\omega)| \doteq 1$, then the effect of the term containing a_{20} will be quite important indeed, since in most practical amplifiers its value then will be quite large compared with that in the normal operating frequency range.

Similar comments apply to the distortion components in the expressions for i_1 given by equations 22 and 37 since one of the terms contained therein appears as a factor in i_3. A frequency invariant term of the same form also appears, however. This is

$$\frac{a_{20}}{Y_{110}(0)F(0)}$$

or

$$\frac{a_{20}}{Z(0)F(0)}$$

depending on the particular device model employed for analysis. Once again, the importance of feedback at direct current on the collector-current component of input signal frequency is noted. In this instance, a large value of $Y_{110}(0)F(0)$ or $Z(0)F(0)$ will make the contribution of these a_{20} terms insignificant compared with the factors containing b_{20} with which they are asssociated. As a critical examination of the distortion terms in equations 22 and 37 shows, possibilities exist for the reduction or complete cancellation of expansion if one utilizes the terms produced in i_1 by second-order distortion products. Since the net amount of compression or expansion now depends on four device coefficients (a_{20}, a_{30}, b_{20}, and b_{30}) and since $Y_{110}(\omega)$ or $Z(\omega)$ is also a factor, the practical achievement of an optimum condition in which the distortion products in i_1 will be a minimum for a number of transistors of the same type is difficult, to say the least. One may use the two equations for i_1 as a guide, however, in designing the circuits in which the transistor is embedded to reduce to a practical minimum the nonlinearity in the input-output relation of the complete amplifier.

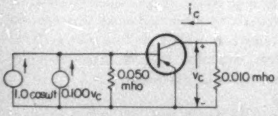


Fig. 10. Transistor output stage of a feedback amplifier for illustrative example

Illustrative Example

The use of some of the equations which have been derived can be illustrated by consideration of a typical example of distortion in the output stage of a transistor feedback amplifier. In so far as signal conditions are concerned, operation may be analyzed using the circuit shown in Fig. 10 (power sources are not shown). It is noted that the combination of elements external to the power transistor is represented by a nonreciprocal network of the form first depicted in Fig. 2. Corresponding to the notation used there, the element values in Fig. 10 are: $y_a=0.500$ mho; $y_b=0$; $y_c=0.0100$ mho; $Y=0.100$ mho; $G_1V_0=1.000$ ampere; $G_2V_0\doteq 0$. It will be assumed that these parameters remain constant throughout the frequency range of interest (including d-c). The transistor linear parameters and distortion coefficients [refer to Fig. 1(A)] are: $g_{11}=0.094$ mho; $g_{12}\doteq 0$; $g_{21}=3.800$ mhos; $g_{22}=0.006$ mho; $a_{20}=-0.200$ ampere per volt^2; $a_{30}=+0.150$ ampere per volt^3; $b_{20}=-1.600$ amperes per volt^2; $b_{30}=1.500$ amperes per volt^3. It is desired to compute the components of the collector circuit current generator, i_0, i_1, i_2, and i_3.

In order to determine the components desired, equations 21 through 24 may be used after A_f is determined utilizing equation 20. Because of the absence of reactive elements and concomitant frequency-dependent transmissions, θ may be considered zero in equation 20 and Y_{110}, T, and F may be treated as constant. The quantities needed in that equation may be found as follows:

$$Y_{110}=g_{11}+y_a=0.144 \text{ mho}$$

$$A_0=\frac{G_0V_0}{Y_{110}}=6.94 \text{ volts}$$

$$T(\omega)=\frac{g_{21}}{Y_{110}}\frac{(Y-g_{12}+y_b)}{(g_{22}+y_c+y_b)}=24.9$$

$$F(\omega)=1+T(\omega)=25.9$$

$$\frac{T(\omega)}{F(\omega)}=0.961$$

$$\frac{A_0}{F}=0.268 \text{ volt}$$

$$\frac{a_{30}}{Y_{110}F}=0.0402$$

$$\frac{b_{30}}{g_{21}}\frac{T}{F}=-0.379$$

$$\frac{a_{20}}{Y_{110}F}=-0.0535$$

$$\frac{b_{20}}{g_{21}}\frac{T}{F}=-0.405$$

Substituting the above quantities in equation 20, the following equation for A_f is obtained:

$$A_f-0.570A_f^3=0.268$$

This equation is satisfied by $A_f=0.281$ volt.

To calculate the various components of current, the following coefficients are required:

$$\left(b_{20}-\frac{g_{21}a_{20}}{Y_{110}}\right)=-1.60-\frac{(3.80)(-0.200)}{0.144}=+3.67$$

$$\left(b_{30}-\frac{g_{21}a_{30}}{Y_{110}}\right)=-1.50-\frac{(3.80)(0.150)}{0.144}=-5.46$$

Now equations 21 through 24 may be used to compute i_0, i_1, i_2, and i_3 as follows:

$$i_0=\frac{(0.281)^2}{2(25.9)}(3.67)=5.60 \text{ ma (milliamperes)}$$

$$i_1=\left\{(3.80)(0.268)+\frac{(0.281)^3}{25.9}\times [(0.750)(-5.46)-(3.67)(-0.405-0.203-0.0535-0.027)]\right\}\cos\omega t$$

$$i_1=\left[1.02+\frac{(0.0222)}{25.9}(-4.10+2.52)\right]\cos\omega t=(1.02-0.00136)\cos\omega t$$

$$i_1=1.02\cos\omega t \text{ amperes}$$

$$i_2=\frac{(0.281)^2(3.67)}{2(25.9)}\cos 2\omega t=5.60\ \cos 2\omega t \text{ ma}$$

$$i_3=[(0.250)(-5.46)-(0.500)(3.67)\times(-0.0535-0.405)]\frac{(0.281)^3}{25.9}\cos 3\omega t$$

$$i_3=(-1.37+0.841)\frac{(0.0222)}{25.9}\cos 3\omega t=-0.454\cos 3\omega t \text{ ma}$$

It is of interest to observe the details of the calculations for i_1 and i_3, particularly since the contributions from the second-order terms (containing b_{20} and a_{20}) have opposite signs to the factors occurring from a_{30} and b_{30}. This effect causes the combined result of the two effects to have a magnitude smaller than each individual contribution.

If one desires the values of the actual currents in the 0.010-mho (10-ohm) load resistor, they may be obtained by dividing the magnitudes shown for i_0, i_1, i_2, and i_3 by $1+g_{22}R_L$ (which is equal to 1.006 in this example).

Conclusions

A study has been made of the amplitude distortion that is produced in a transistor amplifier stage which is part of a feedback amplifier. It has been found that the values of the rectified d-c component, the harmonics, and the fundamental-frequency component of collector current depend in a complicated fashion on several quantities which may be identified with the transistor and its associated active network. The basic factors involved are: the base- and collector-circuit device coefficients (g_{21}, h_{21}, a_{20}, a_{30}, b_{20}, and b_{30}) which in turn depend on the operating point chosen for the transistor; the base-circuit signal level (either base current or base-to-emitter voltage); and the characteristics of the transmission paths from collector circuit to base circuit, both internal and external to the device itself.

The output desired from a transistor amplifier stage, either in terms of power delivered to a prescribed load or peak-to-peak signal over a certain frequency range, is frequently specified. This establishes for the circuit designer both the amount of input signal required and the extent of a rather limited region in which operating points may be selected. The characteristics of the external transmission paths from collector to base are, however, under direct design control, and the impedance presented to the transistor is often under the control of the design engineer as well. The equations presented herein for the signal-frequency and distortion components of collector current provide detailed information regarding the influence of the circuit environment external to the transistor on amplitude distortion. The results, therefore, not only furnish a means of computing distortion in specific circuits, but also provide guidelines for the design of circuitry associated with a transistor stage in which the effects of amplitude distortion are to be controlled.

References

1. Distortion in Feedback Amplifiers, **R. W. Ketchledge.** *Bell System Technical Journal*, New York, N. Y., vol. 34, 1955, pp. 1265–85.

2. Non-Linear Distortion in Transistor Amplifiers at Low Signal Levels and Low Frequencies, **N. I. Meyer.** *Proceedings*, Institution of Electrical Engineers, London, England, vol. 104, pt. C, 1957, pp. 208–16.

3. Network Analysis and Feedback Amplifier Design (book), **H. W. Bode.** D. Van Nostrand Company, Princeton, N. J., 1945, p. 77.

4. Signal Transmission in Non-Reciprocal Systems, **J. H. Mulligan, Jr.** *Proceedings*, Symposium on Active Networks and Feedback Systems, Microwave Research Institute Symposia Series, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., vol. 10, 1961, pp. 125–53.

Transistor Amplifier Stages with Prescribed Gain and Static and Dynamic Sensitivity

J. H. MULLIGAN, JR.
MEMBER AIEE

S. S. SHAMIS
MEMBER AIEE

Synopsis: The many and varied aspects of transistor amplifier sensitivity to changes in temperature and transistor unit replacement have received considerable attention in the literature. Numerous investigations have been reported pertaining to the thermal displacement of the transistor operating point and the variation of its incremental parameters with the shift of the operating point and with temperature.[1–5, 7, 10–16] Furthermore, various schemes for stabilizing the operating point have been considered in detail,[3, 7, 8, 9, 12–16] and design procedures for specific as well as generalized bias network configurations to effect this stabilization have been proposed.[3, 6, 8, 12, 16]

It should be recognized that the design of linear bias circuits is not affected by the origin of the variations against which stabilization must be maintained; the unit-to-unit differences encountered in transistor parameters and static characteristics frequently dwarf the corresponding variations resulting from temperature changes. It will also be noted, that the magnitude and sensitivity of the incremental gain of the amplifier are functions of the bias network parameters as well as those factors which give rise to the need for operating-point stabilization. Clearly, the interdependence of bias stabilization, gain, and gain stabilization would appear to indicate the desirability of an approach to the design problem based upon the simultaneous satisfaction of these, as well as additional, constraints. However, the various treatments accorded the problem in presently available literature do not provide the circuit designer with an orderly procedure for including all of these constraints on the basis noted.

IT IS THE purpose of this paper to develop a procedure for effecting a design to meet prescribed gain and sensitivity criteria in terms of anticipated operational variations and such device data as are readily available from manufacturers' specifications. The amplifier prototype considered herein corresponds to the frequently encountered configuration shown in Fig. 1. The particular problem faced by the designer is that of specifying R_1, R_2, R_C, and R_E, so that the displacement of the operating point, expressed most conveniently as a change in collector current, and the deviation of the incremental gain from a design-center value are held within prescribed limits, despite unit-to-unit variations and temperature effects. Furthermore, the design must be such that any incompatabilities between the gain and stability requirements and the system constraints, i.e., source and load impedances for the amplifier and direct voltage supplies, are resolved.

Theoretical Considerations

Stabilization of Operating Point

As already stated, it is of considerable practical importance to limit the change in collector current that takes place as a result of variation in temperature and

---

Paper 61-96, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 14, 1961; made available for printing December 8, 1961.

J. H. Mulligan, Jr., and S. S. Shamis are with New York University, New York, N. Y.

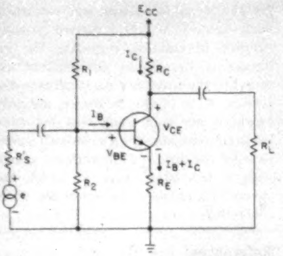


Fig. 1. Common-emitter transistor amplifier with bias network

unit-to-unit differences in transistor-static characteristics. The changes that actually occur depend, of course, not only on the transistor volt-ampere relations but also on the properties of the network in which the device is embedded. It is of interest to consider in detail the properties of the circuit composed of transistor and associated network.

In order to simplify the analysis of the amplifier in Fig. 1, the base-circuit voltage divider, R_1, R_2, is replaced by its Thévenin equivalent, R_B, E_{BB}, in Fig. 2. The elements of the complete amplifier of Fig. 2 which are of interest for d-c conditions are shown in Fig. 3(A). The conclusions to be drawn for this circuit will also apply with slight modification to certain other bias arrangements commonly used. The bias scheme shown in Fig. 3(B), for example, can be shown to be equivalent to that in Fig. 3(A).

The various currents and voltages indicated in Fig. 3(A) can be related by the two Kirchhoff voltage-law equations:

$$E_{BB}=R_EI_C+(R_E+R_B)I_B+V_{BE} \tag{1}$$

$$E_{CC}=(R_E+R_C)I_C+R_EI_B+V_{CE} \tag{2}$$

Naturally, to find a unique solution for I_B and I_C, the transistor characteristics must be introduced by furnishing information concerning the variation of V_{BE} and V_{CE} as functions of I_B and I_C.

Fig. 4 depicts the V_{CE}-I_C family as a function of base current for two different temperatures. It indicates that if base current is held constant, the collector current increases as temperature increases. As a result, if one desired to keep the collector current fixed as the temperature increased, it would be necessary to decrease the base current.

The base-to-emitter volt-ampere characteristics also change with temperature. Representative curves illustrating the variation are shown in Fig. 5. At the temperature T_1 the relation between V_{BE} and I_B can be expressed as

$$V_{BE}=V_{BE_1}+r_1I_B \tag{3}$$

where r_1, the d-c resistance of the forward-biased junction, is a function of I_B. As the temperature changes to $T_2(T_2>T_1)$ there can be written similarly,

$$V_{BE}=V_{BE_1}+\Delta V_{BE}+r_2I_B \tag{4}$$

Although r_1 and r_2 do not differ greatly from one another, a more important consideration here is that their numerical values are usually considerably smaller than the external resistance R_B+R_E in series with the base-emitter junction. It is to be noted particularly that the sign of ΔV_{BE} is negative for a positive temperature change, i.e., V_{BE} decreases as temperature is increased.

To study the problem of stabilizing I_C as temperature is varied and as different transistors of the same type are used, one may treat the characteristics at T_1 and T_2 in Figs. 4 and 5 as representing the extreme variations due to all causes for which suitable compensation is to be provided. In terms of changes in I_C and I_B, the use of equations 3 and 4 in conjunction with 1 provides the expression

$$(R_B+R_E)\Delta I_B+R_E\Delta I_C=-\Delta V_{BE} \tag{5}$$

In formulating this result r_1 and r_2 have been neglected compared with R_B+R_E. The sign of ΔV_{BE} is negative for a positive temperature change. Furthermore, if R_EI_B is small compared with $(R_E+R_C)I_C$, which is usually the case, then a good approximation to equation 2 is

$$E_{CC}=(E_E+R_C)I_C+V_{CE} \tag{6}$$

This is the familiar load-line equation, a typical construction of which is shown in Fig. 4. Since the locus of all possible operating points (V_{CE}-I_C pairs) must be on the load line, the quiescent point and the new operating point, after change in transistor characteristics, must both be located on this line.

Assume now that the desired quiescent point $Q(V_{CE},\ I_C)$, the collector supply voltage, E_{CC}, and the maximum change in collector current, ΔI_C are specified. From this information, as suggested by Fig. 4, one may determine at once the base current (I_{B_1}) needed to provide Q as well as the quantity $\Delta I_B=I_{B2}-I_{B1}$, the change in base current needed to insure that I_C goes to $I_C+\Delta I_C$. It may be determined from Fig. 4 that for small values of positive ΔI_C, resulting in a constant I_C, the required ΔI_B has a negative sign.

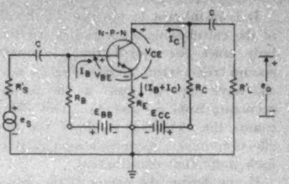


Fig. 2. Transistor amplifier with bias network replaced by Thévenin equivalent

In order that the design objective of ΔI_C be realized, circuit elements external to the transistor must be chosen to insure that the necessary change in I_B does in fact occur. Considering equation 5 it is evident that R_B and R_E must be selected to have values which guarantee that when ΔV_{BE} changes, ΔI_B can assume the proper negative value corresponding to ΔI_C. It is necessary that

$$\left(1+\frac{R_B}{R_E}\right)\Delta I_B=-\frac{\Delta V_{BE}}{R_E}-\Delta I_C \tag{7}$$

Recall that the sign of ΔV_{BE} is negative and that the sign of ΔI_B must also be negative in the problem of direct interest here. On the other hand, the sign of ΔI_C is positive for a positive temperature change. It is clear then that R_E must have a certain minimum value in order that a negative ΔI_B can be obtained. The smallest value of R_E that will insure a given negative ΔI_B in equation 7 is determined by setting $R_B=0$; it is then found that

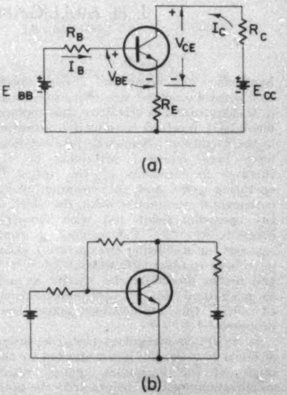


Fig. 3. (A) Portion of transistor amplifier pertinent to d-c analysis. (B) Transistor amplifier with alternative bias network

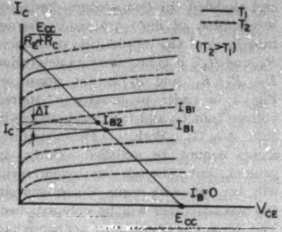


Fig. 4. Shift of collector family as a function of temperature

$$R_{E\min}=-\frac{\Delta V_{BE}}{\Delta I_C+\Delta I_B} \tag{8}$$

For $R_B \neq 0$, the value of R_E necessary to satisfy equation 7 is given by

$$R_E=R_{E\min}\left(1+R_B\frac{\Delta I_B}{\Delta V_{BE}}\right) \tag{9}$$

Merely specifying the quiescent point, ΔI_C, and E_{CC} does not provide sufficient information to determine R_B and R_E uniquely. One may specify sensitivity to a-c-gain changes in addition to the d-c information and solve for R_B and R_E. Certain relations are implied, however, when only the d-c data are given, and it is of interest to examine some of these before turning to the questions of a-c sensitivity and voltage gain. From equation 8 it is clear that the information regarding the operating point and its shift fixes a minimum value of R_E. With the quiescent point specified, the voltage drop across R_E at the initial temperature is $R_{E\min}$ $(I_C+I_{B_1})$. Consequently, the maximum value of R_C that may be used is already fixed at the value

$$R_{C\max}=\frac{E_{CC}-V_{CE}}{I_C}-R_{E\min}\left(1+\frac{I_{B_1}}{I_C}\right) \tag{10}$$

In general, R_C is related to R_E as

$$R_C=R_O-R_E \tag{11}$$

where

$$R_O=\frac{E_{CC}-V_{CE}}{I_C} \tag{12}$$

and I_{B_1}/I_C has been assumed small compared with unity. Since the a-c-load impedance for the amplifier is the parallel combination of R_L' and R_C, neglecting the effect of C_2, specification of operating-point data and d-c shift is seen to have an important effect on the voltage gain that can be realized from the stage. Furthermore, since it is seen from Fig. 2 that R_B shunts the a-c-signal transmission path, it will be impractical to choose $R_B=0$, which resulted in the $R_{E\min}$ given by equation 8. It will be seen in the next section that R_B also affects the sensitivity to gain change. The design problem that must be considered, of course, is the choice of R_B, R_C, and R_E, in order that a suitable compromise may be made among d-c stabilization, a-c sensitivity, and voltage gain.

PROVISION OF A-C SENSITIVITY

In terms of the quantities shown in Figs. 1 and 2, the voltage gain of the amplifier stage is defined as $G_v=e_o/e_s$.

It may be shown that G_v is given approximately by

$$G_v \doteq \frac{R_B}{R_s'+R_B}\times \frac{R_L}{R_E+r_e+(R_s+r_b)\left(\frac{1}{b}+\frac{R_E+r_e+R_L}{r_c}\right)} \tag{13}$$

where

$$R_s=\frac{R_s'R_B}{R_s'+R_B}$$

and

$$R_L=\frac{R_L'R_c}{R_L'+R_c} \tag{14}$$

and r_e, r_b, r_c, and b are the usual small signal transistor parameters. If it is assumed that b changes to $b+\Delta b$ with the other small signal parameters remaining essentially constant in the process, then the fractional change in gain can be written

$$\frac{\Delta G_v}{G_v}\doteq\frac{\Delta b}{b}\,\frac{1}{1+(b+\Delta b)\dfrac{(R_E+r_e)}{R_E+r_e+r_b+R_s}} \tag{15}$$

In the derivation of equations 13 and 15, it has been assumed that $b \gg 1$, $r_e \ll r_c$ $(1-a)$, $r_c \gg r_e+R_E+R_L$, and $r_c \gg r_e+R_E+R_S+r_b$. Equation 15 may be expressed in the form

$$\frac{\Delta G_v}{G_v}=S\frac{\Delta b}{b} \tag{16}$$

where S, the sensitivity function frequently used in feedback theory, is related to the various circuit parameters by

$$\frac{1}{S}=1+(b+\Delta b)\,\frac{(R_E+r_e)}{R_E+r_e+r_b+R_s} \tag{17}$$

As may be seen from equation 16, in order to minimize the percentage change in voltage gain for a prescribed percentage change in b, one should strive to make S as small as possible. If the requirements on R_E imposed by the d-c stabilization conditions as expressed by equations 8 and 9 are combined with equation 17, the sensitivity can be written in terms of R_B as

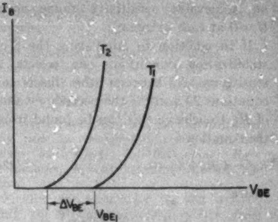


Fig. 5. Shift of emitter-base characteristic as a function of temperature

$$1/S=1+(b+\Delta b)\times \frac{R_{E\min}+kR_B+r_e}{R_{E\min}+kR_B+r_e+r_b+\dfrac{R_s'R_B}{R_s'+R_B}} \tag{18}$$

where

$$1/k=-\left(\frac{\Delta I_C}{\Delta I_B}+1\right) \tag{19}$$

Equation 18 indicates that the sensitivity is a monotonic function of R_B for positive values of R_B. S attains its maximum value when $R_B=0$, and a minimum value, as R_B becomes infinitely large, of $S=1/(1+b+\Delta b)$.

It follows from equations 9 and 11 that R_B cannot be increased indefinitely since, with a prescribed E_{CC}, there is an upper limit for R_E, and hence R_B, beyond which it is not possible to attain the desired quiescent point. The upper limit is evident from equation 11, as being $R_E=R_O$, at which condition G_v has the value zero because $R_C=0$. As a consequence, once the d-c operating point, supply voltage, and shift in collector current, together with the transistor characteristics, are specified, one is limited to a sensitivity S whose values lie between fixed limits. The maximum sensitivity attainable is

$$S_{\max}=\frac{1}{1+\dfrac{(b+\Delta b)(R_{E\min}+r_e)}{R_{E\min}+r_e+r_b}} \tag{20}$$

whereas the minimum is

$$S_{\min}=\frac{1}{1+\dfrac{(b+\Delta b)(R_O+r_e)}{R_O+r_e+r_b+R}} \tag{21}$$

where R is the parallel combination of R_s' and a resistance of value $(R_O-R_{E\min})/k$ ohms. It is to be noted that equations 20 and 21 represent limits of an achievable sensitivity range since $G_v=0$ at each extreme.

If in addition to stipulating the d-c-stabilization conditions one selects a sensitivity S between the limits of equations 20 and 21, the necessary value of R_E to achieve this can be found from the equation

$$R_E^2+A_1R_E+A_0=0 \tag{22}$$

where

$$A_1=kR_s'-R_{E\min}+r_e+\frac{(R_s'+r_b)}{1-\dfrac{S(b+\Delta b)}{1-S}} \tag{23}$$

and

$$A_0=r_e(kR_s'-R_{E\min})+\frac{r_bkR_s'-(R_s'+r_b)R_{E\min}}{1-\dfrac{S(b+\Delta b)}{1-S}} \tag{24}$$

The corresponding value of R_B is then

$$R_B=(R_E-R_{E\min})/k \tag{25}$$

Gain Requirements

Up to this point attention has been concentrated on achieving both d-c and a-c sensitivity. In a practical design, these sensitivities must be achieved simultaneously with a desired G_v. Factors affecting this quantity will now be examined in detail.

If one utilizes equations 11, 14, and 25 in the expression (equation 13) for G_v, with the further assumption that $(R_E+r_e+R_L)/(r_c)\ll 1/b$, the following result is obtained for the voltage gain:

$$G_v=\frac{(R_O-R_E)}{(R_L'+R_O-R_E)}\times \frac{R_L'(R_E-R_{E\min})}{(R_L+r_e+r_b/b)(kR_s'+R_E-R_{E\min})+\dfrac{R_s'}{b}(R_E-R_{E\min})} \tag{26}$$

Although it is desirable to study the variation of G_v with R_E in a direct fashion, a suitable understanding of the variation can be obtained with considerable saving in algebraic manipulation by an indirect approach. The first ratio in this equation is the fraction of the collector-signal current that is supplied to R_L. In the region of interest here, $R_E<R_O$, this factor is seen to decrease monotonically as R_E increases. The second ratio in G_v rises from zero to a maximum and then decreases, finally approaching zero as R_E is continuously increased starting from $R_{E\min}$. This second factor, to be identified as G_{1v}, can be studied conveniently by using $R_E-R_{E\min}$ as the independent variable rather than R_E. It is readily found that the maximum value of G_{1v} occurs when $R_E=R_{E\min}+R_m$ where

$$R_m=\sqrt{kR_s'(r_e+R_{E\min}+r_b/b)} \tag{27}$$

The maximum value of the second factor at this value of R_E is

$$G_{1v\max}=\frac{bR_L'}{R_s'+[\sqrt{bkR_s'}+\sqrt{b(r_e+R_{E\min})+r_b}]^2} \tag{28}$$

Because of the monotonic nature of the first ratio in G_v, the maximum value of G_v will occur at a smaller value of R_E than $R_{\min}+R_m$, and thus it is evident that if $R_O>R_{E\min}+R_m$,

$$G_{v\max}>G=\frac{(R_O-R_{E\min}-R_m)}{(R_L'+R_O-R_{E\min}-R_m)}\times \frac{bR_L'}{R_s'+[\sqrt{bkR_s'}+\sqrt{b(r_e+R_{E\min})+r_b}]^2} \tag{29}$$

Equation 29 is of importance in the amplifier-design procedure because it permits an estimate to be made of the feasibility of achieving a desired G_v while simultaneously having a required sensitivity, S. If the desired G_v is equal to or less than the G given by equation 29, then the value of R_E which will yield the desired G_v may be determined by the use of the equation

$$x^3-B_2x^2-B_1x-B_0=0 \tag{30}$$

in which

$$B_0=kR_s'(R_L'+R_O-R_{E\min})\times(r_e+R_{E\min}+r_b/b) \tag{31}$$

$$B_1=(r_e+R_{E\min}+r_b/b)\times(R_L'+R_O-R_{E\min}-kR_s')+(1+k)R_s'R_L'-(R_O-R_{E\min})\times[(1+k)R_s'+R_L'/G_v] \tag{32}$$

$$B_2=R_L'(1+1/G_v)+R_O-(1+k)R_s'-r_e-2R_{E\min}-r_b/b \tag{33}$$

and the variable x is $R_E-R_{E\min}$. The equation was obtained by solving equation 26 for $R_E-R_{E\min}$.

Design Procedure

Since now a sufficient theoretical background has been established, it is possible to describe a comparatively direct design procedure for determining the unknown elements in the amplifier of Fig. 1. It is helpful to define the design problem again in the light of the results that have been obtained. Given the quiescent operating point V_{CE}, I_C, the permissible shift in collector current, ΔI_C, the collector-supply voltage E_{CC}, the normal and modified (by temperature change or substitution of unit) transistor characteristics, and the terminating resistances R_s' and R_L', it is necessary to determine R_E, R_C, R_1, and R_2 in the circuit of Fig. 1 in order to provide a desired voltage gain G_v and fractional change in gain $\Delta G_v/G_v$. Both of these requirements cannot be satisfied exactly, in general, but instead one requirement will be met with some design margin whereas the other will be only satisfied. The degrees of freedom are not enough to satisfy both exactly. For design convenience, the requirement on fractional gain change is reduced to the specification of a sensitivity S by the use of equation 16.

The design procedure suggested here assigns first priority to achieving, but not exceeding, the permissible shift in collector current, ΔI_C. This position is taken because it is an important practical consideration that effective stabilization of d-c collector current is a first requisite in the control of amplifier performance in the presence of temperature variation and unit-to-unit variations. Starting with E_{CC}, (V_{CE}, I_C), and ΔI_C one can determine ΔI_B from the collector family and ΔV_{BE} from the device data; it is then possible to compute the quantities k, $R_{E\min}$, and R_O using equations 19, 8, and 12. The limits of achievable sensitivity S may then be determined using equations 20 and 21, and an estimate of the voltage gain G_v, that may be attained, can be made using equation 29. When the last three results have been obtained, the designer is able to assess the compatibility of the design requirements. If the desired G_v is less than the G of equation 29 and the desired sensitivity is greater than that computed from equation 21, it is appropriate to proceed to the determination of R_E and the other elements. If these conditions are not met, either the design requirements must be modified or a different type of transistor selected. Some additional compromise in the requirements on G_v and S may still be necessary, however. Knowing that the desired G_v and S may be achieved in the range of values permitted for R_E, the designer may use either equation 22 or 30 to find the exact value of R_E required to satisfy the specification on one quantity. Normally, the first of these two equations is more desirable to use because it is quadratic in contrast to the cubic equation 30. In the range of circuit values commonly encountered, however, equation 30 frequently can well be approximated by reduction to a quadratic equation. If

R_E has been determined on the basis of the requirement on S, using equation 22, it is necessary to check the corresponding value of G_v using equation 26. If the value obtained is less than required, one must decide whether the S or the G_v requirement will take precedence in the design. The final value of R_E is selected after this decision has been made. Finally R_B and R_C are determined using equations 25 and 11, respectively.

In order to specify the elements R_1 and R_2 in Fig. 1, it is necessary to use the values of R_E and R_B determined in equation 1 with values of I_C, I_B, V_{BE}, and V_{CE} at the quiescent point in order to determine E_{BB}. The values of R_1 and R_2 may then be found from the equations

$$R_1 = R_B E_{CC}/E_{BB} \qquad (34)$$

$$R_2 = R_B/(1 - E_{BB}/E_{CC}) \qquad (35)$$

One other aspect of design in connection with the circuit is worth noting. If the quantity E_{CC} is not considered to be fixed, but instead may be treated as a design parameter, then additional design freedom is obtained. If the quiescent operating point is chosen in a region of the collector family where the curves are essentially parallel, then the change in base current, ΔI_B, for a desired ΔI_C can be estimated quite accurately without benefit of the load line, which is established as soon as V_{CE}, I_C and E_{CC} are specified. If E_{CC} is not specified, then R_O as given by equation 12 is not fixed and may be treated as an additional design parameter. In particular, if there is difficulty in meeting the gain requirement, it may be seen from equation 29 that a sufficiently large increase in R_O can cause G to approach the G_{1v} given by equation 28. In addition, if a sufficiently large R_O can be achieved, the minimum sensitivity given by equation 21 can be made to approach the value wherein R is replaced by R_S'.

Design examples in the next section illustrate some of the procedures discussed.

Illustrative Examples

The use of the design techniques discussed in this paper is readily demonstrated by means of specific examples.

1. An amplifier is to be designed to operate between a source impedance of 200 ohms and a load impedance of 2,500 ohms in a system provided with a 35-volt d-c power supply. The voltage gain is required to be not less than 3 and must be held constant within 10% of the value ultimately selected for nominal gain. The transistor employed in the design is operated nominally at a quiescent point such that $V_{CE}=7$ volts, $I_C=7$ ma (milliampere), $V_{BE}=0.9$ volts, $I_B=0.4$ ma, and it is stipulated that the collector-current variation must be limited to 2 ma.

The collector characteristics and device data indicate that an estimate of 0.25 ma for ΔI_B is appropriate and that a ΔV_{BE} of 0.2 volt should be allowed for. The small signal-transistor parameters are $r_e=5$ ohms, $r_b=200$ ohms, and the range of b to be tolerated is 20–80 (so that $b = 20$, $\Delta b = 60$). Since the $\Delta G_v/G_v$ must be less than 10% and since a $\Delta b/b$ of 3 must be tolerated, the sensitivity required is 3.3%. The design proceeds as follows:

From equation 19, $k=1/7$
From equation 8, $R_{E\min}=114$ ohms
From equation 12, $R_O=3{,}710$ ohms
From equation 20, $S_{\max}=3.25\%$
From equation 21, $S_{\min}=1.37\%$
From equation 27, $R_m=60.8$ ohms, which is less than $R_O-R_{E\min}$
From equation 29, $G_{v\max}>5.06$

Clearly, the sensitivity and gain requirements of the amplifier lie within the limitations noted above, so that the design may be completed. Since the $\Delta G/G$ required is less stringent than that which may be realized for $R_E=R_{E\min}$, specifying a design value of $S = 3\%$ is not unreasonable, whence:

From equation 22, $R_E=242$ ohms
From equation 25, $R_B=426$ ohms
From equation 11, $R_C=3{,}468$ ohms
From equation 1, $E_{BB}=2.66$ volts
From equation 34, $R_1=5{,}600$ ohms
From equation 35, $R_2=462$ ohms

The resistors to be used in the amplifier may now be designed, using standard values, as follows: $R_E=240$ ohms, $R_C=3{,}300$ ohms, $R_1=5{,}600$ ohms, $R_2=470$ ohms. Finally, the values for G_v and $\Delta G_v/G_v$ using the designated resistors are, from equations 13 and 15, respectively: $G_v=3.7$, $\Delta G_v/G_v=8.65\%$ (or $S=2.88\%$).

2. Suppose that the sensitivity requirement for the amplifier discussed above is to be changed so that $S=2\%$, with the remaining specifications to be met as stated. If the design procedure is followed as previously, it is found that

From equation 22, $R_E=605$ ohms
From equation 25, $R_B=3{,}437$ ohms
From equation 11, $R_C=3{,}105$ ohms

On the basis of these values, it is found from equation 26 that the resulting gain is only 2.07, as compared with the amplifier specification of 3. Since k, $R_{E\min}$, and R_E are fixed by operating point and gain sensitivity requirements, the design freedom obtained by treating R_O as an additional parameter becomes strikingly apparent from an examination of equation 26 for the values of k, $R_{E\min}$, and R_E applicable herein:

$$G_v = \frac{R_O - 605}{R_O + 1{,}895}\,(3.75)$$

It is readily established from the above relation that a voltage gain of 3 may be obtained for an R_O of 10,615 ohms; this requires, from equation 12, a d-c supply voltage $E_{CC}=83$ volts. If this supply-voltage requirement can be accommodated, the design may proceed as in the previous example.

From equation 11, $R_C=10{,}010$ ohms
From equation 1, $E_{BB}=6.55$ volts
From equation 34, $R_1=43.5$ kilohms
From 35, $R_2=3{,}730$ ohms

The resistors to be used in the amplifier may now be designated, using standard values, as follows: $R_E=620$ ohms, $R_C=10$ kilohms, $R_1=43$ kilohms, $R_2=3.6$ kilohms.

Finally, the values for G_v and $\Delta G_v/G_v$ using the designated resistors are from equations 13 and 15, respectively: $G_v=2.95$, $\Delta G_v/G_v=5.96\%$ (or $S=1.99\%$).

Conclusions

The problem of designing bias circuits for transistor amplifiers is one of resolving conflicting requirements of operating-point stability, gain, and gain sensitivity, while taking into account specific system constraints and substantial variations in the device characteristic. The design procedure developed in this paper permits the solution of the amplifier-design problem by means of an orderly sequence of calculations. The procedure also indicates readily when the given performance requirements are mutually incompatible, i.e., when a solution to the design problem cannot be obtained in the presence of the constraints imposed.

The results of the analysis also indicate that additional design facility may be acquired by allowing a particular factor normally considered to be a system constraint, such as the supply voltage, to be treated as an additional design parameter. The possibility of utilizing the additional design freedom becomes strikingly apparent as one employs the design procedure presented.

References

1. VARIATION OF TRANSISTOR PARAMETERS WITH TEMPERATURE, **C. R. Eshelman.** *Semiconductor Products*, New York, N. Y., Jan./Feb., 1958.

2. TEMPERATURE DEPENDENCE OF JUNCTION PARAMETERS, **W. W. Gaertner.** *Proceedings*, Institute of Radio Engineers, New York, N. Y. vol. 45, May 1957.

3. BIAS CONSIDERATIONS IN TRANSISTOR CIRCUIT DESIGN, **S. K. Ghandhi.** *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, vol. CT-4, Sept. 1957.

4. POWER TRANSISTORS FOR AUDIO OUTPUT CIRCUITS, **L. J. Giacoletto.** *Electronics*, New York, N. Y., Jan. 1954.

5. INVESTIGATION OF POWER GAIN AND TRANSISTOR PARAMETERS AS FUNCTIONS OF TEMPERATURE, AND FREQUENCY, **A. B. Glenn, I. Joffe.** *Convention Record*, Institute of Radio Engineers, vol. IV, Part 3, "Electron Devices and Receivers," 1956.

6. A GENERALIZED THEORY OF TRANSISTOR BIAS CIRCUITS, **H. Hellerman.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 76, 1957, pp. 694–97.

7. TEMPERATURE EFFECTS IN CIRCUITS USING JUNCTION TRANSISTORS (book), **H. C. Lin, A. A. Barco.** Transistors I, RCA Laboratories, Princeton, N. J., 1956.

8. Transistor Bias Stabilization, P. Penfield. *Audio*, Mineola, N. Y., July 1956.

9. Transistor Operation: Stabilization of Operating Points, R. F. Shea. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 40, Nov. 1952.

10. The Variation of the Forward Characteristics of Junction Diodes With Temperature, R. F. Shea, J. S. Schaffner. *Ibid.*, vol. 43, 1955.

11. Temperature Stabilization of Transistor Amplifiers, L. M. Vallese. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, Sept. 1956, pp. 379–84.

12. Transistor Bias Design From Thermal Incremental Properties, L. M. Vallese. *Electronic Engineering*, London, England, Feb. 1959.

13. Transistor Electronics (book), D. Dewitt, A. L. Rossoff. McGraw-Hill Book Company, Inc., New York, N. Y., 1957.

14. Transistor Electronics (book), A. W. Lo, R. O. Endres, J. Zawels, F. D. Waldhauer, C. Cheng. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1955.

15. Principles of Transistor Circuits (book), R. F. Shea. John Wiley & Sons, Inc., New York, N. Y., 1953.

16. Transistor Circuit Engineering (book), R. F. Shea. John Wiley & Sons, Inc., New York, N. Y., 1957.

The Charging of Nonspherical Particles in a Corona Discharge

P. L. SMITH
ASSOCIATE MEMBER AIEE

G. W. PENNEY
FELLOW AIEE

ELECTROSTATIC precipitation of dust may be considered as occurring in two stages: the charging stage, in which the dust particles are charged by an ion stream, and the collecting stage, in which the charged particles are deposited on one electrode by an electric field. The two stages proceed simultaneously in the industrial-type precipitator. The efficiency of the collecting process depends among other things on the effectiveness of the charging process: the particle charge should be as large as possible.

Pauthenier and Moreau-Hanot[1] derived equations for the charging of spherical dust particles, and measurements on perfect spheres of diameters greater than one micron have confirmed their theory. However, two observers, P. Cooperman (in communication with the authors) and J. G. Hewitt, Jr.,[2] have noted that for particles which are not quite spherical (*BB* shot, in particular) the measured particle charge is noticeably higher than that predicted from the theory. Fig. 1, adapted from Hewitt's thesis, is typical of these observations. Extrapolating the observations, one might expect that particles of markedly nonspherical shape could attain charges much higher than would be predicted from the spherical-particle theory. Since actual dust particles are not all true spheres, the implications of this are clear: if the particle charge is higher than predicted by the theory, the precipitation process will be favorably affected. A knowledge of the magnitude of this effect would be valuable in precipitator design. The present investigation comprises an analytical study of the charging of ellipsoidal particles and an experimental study of the charging of particles of various other shapes.

Charging of Spherical Particles

In electrostatic precipitation, the corona discharge is used as a source of ions of one sign. A particle of the dust to be precipitated acquires an electric charge by capturing ions which approach its surface. Some ions approach the particle surface by moving along lines of the electric field which terminate on the particle; others are brought into the vicinity by their random motion. For particles larger than one micron in diameter, motion along the electric field lines is the more important; it is these "large" particles with which this paper is concerned. The particles are assumed to have sufficient conductivity that they may be considered conductors in the electrostatic sense.

The work of Pauthenier and Mme. Moreau-Hanot[1] forms the basis for the present-day theory of the charging of large particles in a corona discharge. They evaluated the charging of a spherical conducting particle placed in a uniform electric field, with a uniform density of unipolar ions moving through the field. The ion current to the particle is equated to the rate of change of the particle charge q, ~~and the solution to~~ the resulting differential equation indicates that the charge on the particle approaches a limiting or saturation value given by

$$q_s = 12\pi\epsilon E_0 a^2 \qquad (1)$$

(Rationalized meter-kilogram-second units are used in all equations.)

The time required to attain this charge is usually of the order of a few milliseconds; because this time is short compared to the total time that the particle spends in the charging region (at least in the single-stage units), the particle is often considered to acquire its saturation charge instantaneously. This attitude has been adopted here, and in the work on nonspherical particles only the saturation charge has been determined. The determination of the charging time is much more difficult for nonspherical particles, and the expected usefulness of the results was not sufficient to justify undertaking the problem.

Of particular interest is the quantity, obtained by dividing the saturation charge by the surface area of the particle, which will be called the charge/area ratio; for a spherical particle, the charge/area ratio is

$$\frac{q_s}{A_s} = \frac{12\pi\epsilon E_0 a^2}{4\pi a^2} = 3\epsilon E_0 \qquad (2)$$

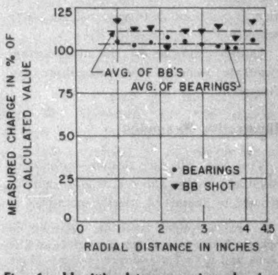


Fig. 1. Hewitt's data comparing charging of perfect (ball bearings) and imperfect (BB shot) spheres in corona from a 6.3-mil wire inside a concentric cylinder 9 inches in diameter at $V_w = +48$ kv

---

Paper 61-93, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 27, 1960; made available for printing December 13, 1960.

P. L. Smith and G. W. Penney are with the Carnegie Institute of Technology, Pittsburgh, Pa.

This investigation was supported by Aerotec Industries, Inc. The authors wish to acknowledge the suggestions and encouragement of E. R. H. Pegg and S. Orem, Jr., of the Aerotec firm. The work was done in part by P. L. Smith in partial fulfillment of the requirements for a Ph.D. degree at Carnegie Institute of Technology.

The charge/area ratio can be shown by dimensional analysis to depend on the shape of the particle but not upon its size; therefore this quantity will form a useful basis for comparing the charging of particles of various shapes. Such comparisons should, however, be made with caution, since in some cases (for instance with particles having re-entrant cavities in their surfaces) the charge/area ratio determined as above may have little relevance to the precipitation properties of these particles.

That the particles to be precipitated are frequently not spherical can be readily determined by microscopic examination. Thus it will be of interest to determine whether the particle shape has any appreciable effect in the electrostatic precipitation process. The measurements made by Cooperman and Hewitt using *BB* shot as test particles indicate that the effect on the charging process in particular may well be significant.

Charging of Ellipsoidal and Spheroidal Particles

In general, the charging of an irregular particle shape cannot be treated analytically. However, useful information can be obtained by studying shapes which are susceptible to analytical treatment, and which approximate the shapes of some real particles. This category includes ellipsoidal particles, which in turn possess as limiting cases spheroidal, spherical, discoidal, and rodlike particles.

The saturation charge acquired by an ellipsoidal conducting particle will be calculated by a method somewhat simpler than the method of Pauthenier and Moreau-Hanot, but which takes no account of the charging time. This method, discussed more fully elsewhere,[3] makes use of the fact that when the surface charge at every point on the surface of a particle has the same sign as the incident ions (or zero), the ions will be repelled away and the particle can acquire no further charge. The surface charge density is evaluated as the sum of the charge density induced on the surface of the uncharged particle when it is placed in the charging field and the density of the free charge q which the particle has acquired.

The induced surface charge density is calculated for an ellipsoidal particle in the Appendix, the result being

$$\sigma_1 = \frac{-2\epsilon E_0 \sqrt{\dfrac{(\eta+a^2)(\zeta+a^2)}{(b^2-a^2)(c^2-a^2)}}}{a\sqrt{\eta\zeta}\displaystyle\int_0^\infty \frac{d\xi}{(\xi+a^2)R_\xi}} \qquad (3)$$

when the conducting ellipsoid is oriented with its long axis a parallel to the charging field E_0.

The distribution of the free charge is given by[4]

$$\sigma_2 = \frac{q}{4\pi\sqrt{\eta\zeta}} \qquad (4)$$

Since each of these distributions has the surface of the conducting ellipsoid as an equipotential, they may be superposed and their sum will have the boundary of the particle as an equipotential, in accordance with the requirements of the problem.

If the sum $\sigma_1+\sigma_2$ is to have the same sign as the incident ions at all points on the surface of the particle, the value of q must satisfy the condition

$$q \geq \frac{8\pi\epsilon E_0 \sqrt{\dfrac{(\eta+a^2)(\zeta+a^2)}{(b^2-a^2)(c^2-a^2)}}}{a\displaystyle\int_0^\infty \frac{d\xi}{(\xi+a^2)R_\xi}} \qquad (5)$$

for all values of η and ζ. The maximum values of η and ζ are $-c^2$ and $-b^2$ respectively[5]; the inequality of equation 5 is therefore satisfied if

$$q = q_s = \frac{8\pi\epsilon E_0}{a\displaystyle\int_0^\infty \frac{d\xi}{(\xi+a^2)R_\xi}} \qquad (6)$$

The integral in equation 6 is an elliptic integral which may be evaluated in terms of tabulated functions.[6,7] However, to calculate the charge/area ratio for the ellipsoidal particles, the surface area of an ellipsoid must be computed; this leads to further elliptic integrals[8] and the final result cannot be interpreted without rather tedious numerical computations. Therefore, the saturation charge of equation 6 has been evaluated only for the limiting cases of the prolate and oblate spheroids, for which both the integral appearing there and the surface area of the particle can be expressed in terms of elementary functions.

In the special case $a>b=c$, the ellipsoid becomes a prolate spheroid; for this case the saturation charge has the value

$$q_s = 4\pi\epsilon E_0 b^2 \times \left[\frac{\left[\left(\dfrac{a}{b}\right)^2-1\right]^{3/2}}{\left(\dfrac{a}{b}\right)\ln\left(\dfrac{a}{b}+\sqrt{\left(\dfrac{a}{b}\right)^2-1}\right)-\sqrt{\left(\dfrac{a}{b}\right)^2-1}}\right] \qquad (7)$$

when the long axis a of the spheroid is parallel to the charging field. The charge/area ratio, obtained by dividing

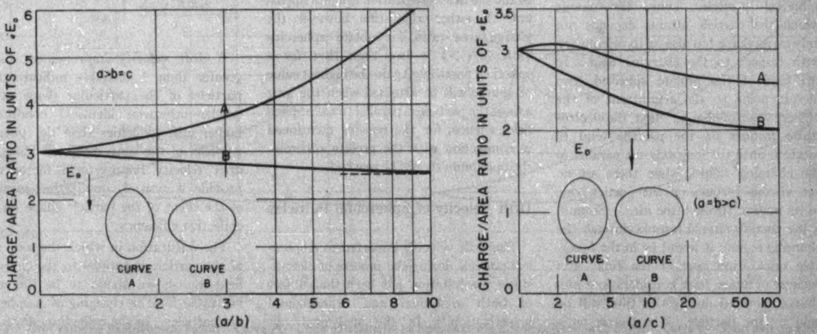


Fig. 2. Charge/area ratio for prolate and oblate spheroids. Curve A—Long axis parallel to charging field E_0. Curve B—Short axis parallel to E_0

Left—For prolate spheroid
Right—For oblate spheroids

this saturation charge by the surface area of the prolate spheroid,[9] is

$$\left.\frac{q_s}{A_{pr}}\right]_{a\parallel E_0} = \frac{2\epsilon E_0\left[\left(\frac{a}{b}\right)^2-1\right]^2}{\left[\left(\frac{a}{b}\right)\ln\left(\frac{a}{b}+\sqrt{\left(\frac{a}{b}\right)^2-1}\right)-\sqrt{\left(\frac{a}{b}\right)^2-1}\right]\left[\sqrt{\left(\frac{a}{b}\right)^2-1}+\left(\frac{a}{b}\right)^2\tan^{-1}\sqrt{\left(\frac{a}{b}\right)^2-1}\right]} \tag{8}$$

This equation shows that the charge/area ratio does not depend on the size of the spheroid (i.e., on a or b) but only on its shape (i.e., on the axial ratio a/b).

The dependence of the charge/area ratio for prolate conducting spheroids with this orientation on the axial ratio (a/b) is illustrated in Fig. 2 (left). The elongation of the particle produces an increased value of the charge/area ratio; as the particle becomes rodlike $(a/b \to \infty)$ the value becomes quite large. However, the charge/area ratio is only twice as great for an axial ratio $a/b=9$ as for a spherical particle; thus, to produce a sizable increase in charge, the deviation from spherical shape must be considerable.

When the prolate spheroid is oriented with its short axis c parallel to the charging field, the charge/area ratio is found in a similar fashion to be

$$\left.\frac{q_s}{A_{pr}}\right]_{c\parallel E_0} = \frac{4\epsilon E_0\left[\left(\frac{a}{b}\right)^2-1\right]^2}{\left[\left(\frac{a}{b}\right)\sqrt{\left(\frac{a}{b}\right)^2-1}-\ln\left(\frac{a}{b}+\sqrt{\left(\frac{a}{b}\right)^2-1}\right)\right]\left[\sqrt{\left(\frac{a}{b}\right)^2-1}+\left(\frac{a}{b}\right)\tan^{-1}\sqrt{\left(\frac{a}{b}\right)^2-1}\right]} \tag{9}$$

The dependence of this quantity on the axial ratio (a/b) is also illustrated in Fig. 2 (left). In this orientation the charge/area ratio is less than the value for a spherical particle. Thus, the charge which the particle attains depends not only on its shape but also on its orientation with respect to the charging field. In the case of the prolate spheroid, two factors point to the attainment of the highest value of charge: first, the electrostatic torques on the particle tend to rotate it until the long axis a is parallel to the charging field[10], while there are no net viscous torques on the particle due to its motion through the air.[11] Second, if the particle should tumble through the charging region, it would be in the favorable orientation part of the time, and because of image force considerations any charge acquired during this time will be held by the particle. The charge on a prolate spheroidal particle will therefore approach the value that would be attained if the particle were held fixed with its long axis a parallel to the charging field.

When $a=b>c$, the ellipsoid reduces to an oblate spheroid. The expressions for the charge/area ratio in this case are found to be: with the long axis a parallel to the charging field

$$\left.\frac{q_s}{A_{ob}}\right]_{a\parallel E_0} = \frac{4\epsilon E_0\left[\left(\frac{a}{c}\right)^2-1\right]^2}{\left[-\sqrt{\left(\frac{a}{c}\right)^2-1}+\left(\frac{a}{c}\right)^2\tan^{-1}\sqrt{\left(\frac{a}{c}\right)^2-1}\right]\left[\left(\frac{a}{c}\right)\sqrt{\left(\frac{a}{c}\right)^2-1}+\ln\left(\frac{a}{c}+\sqrt{\left(\frac{a}{c}\right)^2-1}\right)\right]} \tag{10}$$

and with the short axis c parallel to the charging field

$$\left.\frac{q_s}{A_{ob}}\right]_{c\parallel E_0} = \frac{2\epsilon E_0\left[\left(\frac{a}{c}\right)^2-1\right]^2}{\left(\frac{a}{c}\right)\left[\sqrt{\left(\frac{a}{c}\right)^2-1}-\tan^{-1}\sqrt{\left(\frac{a}{c}\right)^2-1}\right]\left[\left(\frac{a}{c}\right)\sqrt{\left(\frac{a}{c}\right)^2-1}+\ln\left(\frac{a}{c}+\sqrt{\left(\frac{a}{c}\right)^2-1}\right)\right]} \tag{11}$$

The dependence of these quantities on the axial ratio (a/c) is shown in Fig. 2 (right). The effect on the charge of the particles orientation is again apparent. In either orientation, however, the charge/area ratio for oblate spheroids with $(a/c)>4$ is less than that for a spherical particle. Again the highest value of charge will be attained when the long axis of the particle is parallel to the applied field; hence, for the reasons mentioned in connection with the prolate spheroid, the maximum should be reached.

Drift Velocity of Spheroidal Particles

The drift velocity imparted to a spherical particle during the process of electrostatic precipitation has been the subject of both analytical and experimental investigation.[12] In the analytical calculation of the drift velocity, the viscous resistance of the medium to the motion of the particle is assumed to follow Stokes' law; then the drift velocity may be determined from the condition for equilibrium between the coulomb force on the charged particle and the viscous drag opposing its motion toward the collecting electrode.

Since analytical expressions for the viscous resistance of spheroidal particles are available,[13] it may be of interest to compute the drift velocities of the spheroids. As is the case for the spherical particles, the drift velocities of spheroidal particles are directly proportional to the particle size (aside from the dependence on particle shape). To stress particularly the effect of particle shape on drift velocity, a "drift velocity improvement factor" has been calculated by dividing the drift velocity of the spheroidal particle by that of a spherical particle having the same weight as the spheroid; the radius of a spherical particle having the same weight as a spheroid of semiaxes a, b, c is $\sqrt[3]{abc}$, and its drift velocity is[12]

$$V_s = \frac{2\epsilon E_0 E_c \sqrt[3]{abc}}{\lambda} \tag{12}$$

A drift velocity improvement factor greater than 1 therefore indicates that particles of the particular shape attain (under otherwise identical conditions) higher drift velocities than do spherical particles of the same weight. Thus, the drift velocity improvement factor may provide a partial, qualitative measure of the effect of the particle shape on the collection efficiency.

The orientation in which the long axis of the particle is parallel to the charging field was shown above to be the most favorable for the charging of a spheroidal particle. In the collecting stage, the long axis of the particle will again tend to line up parallel to the field; since the particle drifts in the direction of the collecting field, its motion will then be in a direction parallel to its long axis. For a spheroidal particle this is also the direction of least viscous resistance to motion through the fluid.

For prolate spheroids charged and collected with the long axis parallel to the respective fields, the drift velocity improvement factor is

$$\text{DVIF}\Big]_{\substack{\text{prolate spheroid}\\ a\parallel E_0,\ E_c}}=\frac{\left[2\left(\frac{a}{b}\right)^2-1\right]\cosh^{-1}\left(\frac{a}{b}\right)-\left(\frac{a}{b}\right)\sqrt{\left(\frac{a}{b}\right)^2-1}}{4\sqrt[3]{\frac{a}{b}}\left[\left(\frac{a}{b}\right)\ln\left(\frac{a}{b}+\sqrt{\left(\frac{a}{b}\right)^2-1}\right)-\sqrt{\left(\frac{a}{b}\right)^2-1}\right]}\tag{13}$$

A plot of this quantity as a function of the axial ratio (a/b) is shown in Fig. 3 (left). In contrast, the opposite extreme, with both charging and collection taking place with the particle in the least favorable orientation (that with its short axis parallel to the fields) gives a drift velocity improvement factor of

$$\text{DVIF}\Big]_{\substack{\text{prolate spheroid}\\ c\parallel E_0,\ E_c}}=\frac{\left[2\left(\frac{a}{b}\right)^2-3\right]\cosh^{-1}\left(\frac{a}{b}\right)+\left(\frac{a}{b}\right)\sqrt{\left(\frac{a}{b}\right)^2-1}}{4\sqrt[3]{\frac{a}{b}}\left[\left(\frac{a}{b}\right)\sqrt{\left(\frac{a}{b}\right)^2-1}-\ln\left(\frac{a}{b}+\sqrt{\left(\frac{a}{b}\right)^2-1}\right)\right]}\tag{14}$$

The values of this quantity have also been shown in Fig. 3 (left) for comparative purposes. In each case the particle is assumed to acquire its saturation charge while in the charging stage. The upper curve in Fig. 3 (left) represents the situation most likely to occur in actual precipitation, since the orientation with the long axis parallel to the fields is preferred by the particle. On this basis, the drift velocity for prolate spheroids would be expected to be greater than that for spherical particles of the same weight.

A similar computation gives for the oblate spheroids in the favorable orientation with the long axis parallel to the fields a drift velocity improvement factor of

$$\text{DVIF}\Big]_{\substack{\text{oblate spheroid}\\ a\parallel E_0,\ E_c}}=\frac{\sqrt[3]{\frac{a}{c}}\left\{\left[3\left(\frac{a}{c}\right)^2-2\right]\cos^{-1}\left(\frac{c}{a}\right)-\sqrt{\left(\frac{a}{c}\right)^2-1}\right\}}{4\left[-\sqrt{\left(\frac{a}{c}\right)^2-1}+\left(\frac{a}{c}\right)^2\tan^{-1}\sqrt{\left(\frac{a}{c}\right)^2-1}\right]}\tag{15}$$

In the unfavorable orientation with the short axis c parallel to the field, the value is

$$\text{DVIF}\Big]_{\substack{\text{oblate spheroid}\\ c\parallel E_0,\ E_c}}=\frac{\left[\left(\frac{a}{c}\right)^2-2\right]\cos^{-1}\left(\frac{c}{a}\right)+\sqrt{\left(\frac{a}{c}\right)^2-1}}{4\sqrt[3]{\left(\frac{a}{c}\right)^2}\left[\sqrt{\left(\frac{a}{c}\right)^2-1}-\tan^{-1}\sqrt{\left(\frac{a}{c}\right)^2-1}\right]}\tag{16}$$

The values of these quantities are plotted in Fig. 3 (right). Again it is evident that a higher drift velocity will be attained by the spheroidal particles than by spherical particles of the same weight.

To the extent, then, that the drift velocity of the particle determines the precipitation efficiency, the efficiency of collection of spheroidal particles will be greater on a weight basis than will the efficiency of collection of spherical particles of the same particle-weight distribution.

Experimental Apparatus

The preceding analysis indicates that although a nonspherical particle may acquire a higher charge than a perfect sphere, the shape must deviate considerably from the spherical before a sizable increase is observed. Since this conclusion is at variance with what one would normally infer from the measurements made using *BB* shots, an experimental study of particle charging was undertaken. The object was to confirm these prior observations and to study particle shapes which can not be treated analytically. The ball-dropping method[14-16] was chosen as the basic experimental technique, although modifications were required for the present work.

Basically, the ball-dropping method involves dropping or shooting relatively large particles (3/16-inch-diameter steel ball bearings have been used) through the corona cylinder. The particles are caught in a Faraday cage where the charge acquired is measured. The measured value of charge may then be compared

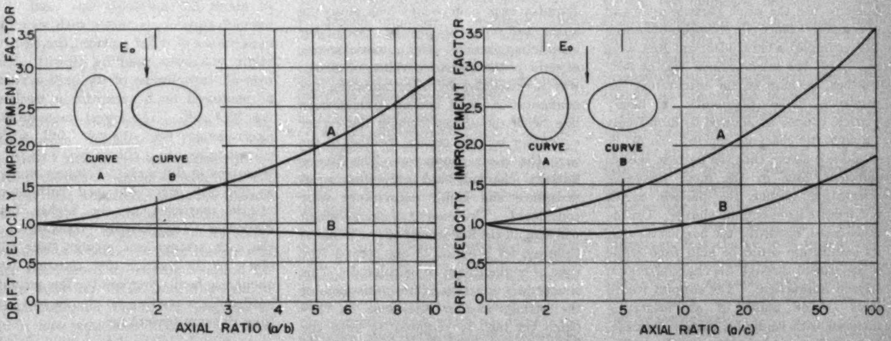


Fig. 3. Drift velocity improvement factor. Curve A—Long axis parallel to both charging E_0 and collecting E_c fields. Curve B—Short axis parallel to E_0 and E_c

Left—For prolate spheroids
Right—For oblate spheroids

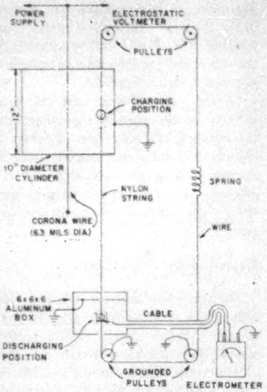


Fig. 4. Experimental apparatus

with the value calculated from equation 1. Concentric cylinder electrodes are usually used, since the field strength in the charging region can be calculated for this configuration. Cooperman's measurements[14] show a scatter of some 10% about the calculated values of particle charge. Hewitt's data[2,16] show less scatter, but all his measured values of charge are higher than the theoretical values, typically by some 5% to 15%. Both observers averaged the readings for several particles in order to obtain sufficient sensitivity.

If the readings must be averaged, it is apparent that, in the case of nonspherical particles, either several identical particles must be made, or arrangements must be made to use the same particle over and over again. Since the prospects of finding (for example) several identical *BB*s were not good, the latter course was chosen. Further, because of the orientation dependence noted previously, a heavy particle cannot be allowed to drop freely through the charging region. To do so would not assure that the particle spends sufficient time in the most favorable orientation; hence the particle might not acquire its maximum charge. On the other hand, mounting the particle so that its orientation could be held fixed while it was passed through the charging region seemed impractical. The support would have to be made of an insulating material with negligible surface leakage; it would have to hold the particle rigidly, yet not appreciably distort the field around the particle.

The foregoing requirements were compromised by drilling a small hole, 0.030 inch in diameter, through the particle and passing a piece of nylon fishing line through the hole. The particle was glued to the line and the line was then looped over a system of pulleys, thus permitting the particle to be moved slowly through the charging region. The final form of the apparatus is shown in Fig. 4. Because of twist in the string, the particle rotated about its axis (the string) as it was moved up and down; nevertheless, it spent sufficient time in any given orientation to acquire substantially its maximum charge. The line was spring-loaded to a tension of about one pound; since the electrostatic force on a fully charged 3/16-inch sphere is less than 10^{-4} pound, this tension prevented the electrostatic force from deflecting the particle toward the outer cylinder.

The string itself acquires a charge when in the corona region and, because it is an insulator, will hold this charge as the string passes through the Faraday cage. This charge induces a charge distribution on the cage, and the needle of the electrometer is observed to fluctuate wildly as the string is moved up and down. To make readings possible, a section of the string was replaced with a wire which passed over a grounded pulley. With the particle in the charging position, the wire section was inside the Faraday cage, and a reading not influenced by the charge on the string could be taken. This scheme was checked for hysteresis effects by moving the particle down until it was almost, but not quite, in contact with the cage and then moving it out again. The cage voltage always returned to zero when the particle was returned to the up position.

While the particle is in contact with the Faraday cage, charge may leak along the string and onto the cage by way of the conducting particle. This potential source of error in the measured values of charge was investigated by approximating the configuration as a coaxial transmission line, with the string forming the center conductor and the central member of the cage the outer conductor. The line is initially charged, and only the series resistance and shunt capacitance were considered. The amount of charge which will leak off the line when one end is short-circuited for a period of one second (the time of contact) can be calculated. This charge is a function of the resistance of the string, and a resistance of 10^{18} ohms per inch is required to keep the leakage charge less than 1.0% of the charge to be measured. A resistance of this order of magnitude was obtained by first drying the string in an oven at 50 degrees centigrade for 2 hours. After attaching the particle, the string was coated for several inches on either side of the particle with ceresin wax. In spite of these precautions, the surface leakage was excessive when the relative humidity in the laboratory was greater than about 40%, and all the data reported here were taken when the relative humidity was below 40%.

Since the particle was to be moved slowly through the charging region, the corona cylinder itself was made quite short, as may be seen in Fig. 4. The region of truly cylindrical field may be a band only an inch or so high midway along the length of the cylinder. However, the particle may be moved through this region as slowly as required to attain the saturation charge. In the experiments conducted it appears that hand operation allowed sufficient time for the maximum charge to be attained.

The central member of the Faraday cage itself, where the particle is discharged, was designed to assure that the particle will be completely discharged upon contact. This was accomplished in part by putting most of the capacitance of the system in the electrometer and cable: 30 $\mu\mu$f (micromicrofarads) out of a total of 35.4 $\mu\mu$f. However, to assure complete discharging, the particle was surrounded as completely as was feasible with metal at the same potential. Then, when the particle makes contact, the charge will migrate to the outer surface of the central member, away from the particle.

To obtain high sensitivity, the capacitance of the system must be as low as possible. Accordingly, a low-capacitance cable, type *RG-114/U* (with a capacitance of about 6.5 $\mu\mu$f/foot) was used. A vacuum-tube electrometer with an input capacitance of 6 $\mu\mu$f replaced the electrostatic voltmeter used by Hewitt. The over-all capacitance of the cage system, as measured by a substitution method, was 35.4 $\mu\mu$f. The lowest range on the electrometer was 0.8-volt full scale, corresponding to a charge of 2.84×10^{-11} coulombs at full scale. Voltages as high as 20 volts were obtained with some of the particles at the higher field strengths. The leakage resistance of the cage system was greater than 10^{14} ohms, so the leakage time constant was about one hour; hence in the few seconds needed for a reading no charge was lost from the cage by leakage.

Experimental Procedure

To measure the charge acquired by a particular particle, the particle was mounted as described above and placed in the corona region. With the particle in the charging position (Fig. 4), the Faraday cage was discharged by means of a short-circuiting switch in the electrometer. The particle was then moved down until it contacted (and gave up its charge to) the central member of the cage. After moving the particle back to the charging position, the voltage on the cage was read from the electrometer.

The particle was in contact with the cage for about one second. Periodic checks for leakage charge were made by leaving the particle in contact with the cage for one minute. If the voltage on the cage was then found to be more than about 2.0% higher than for the 1-second contact, the measurements were discarded. As previously noted, leakage was significant when the relative humidity was greater than 40%.

Reproducibility was checked by repeating the measurements several times, and by further trials on different days. In no case did any of the measurements fall outside a range of ±3.0% from the average value.

The equations used to calculate the field strength in the corona region are:

$$E(r)=E_l\sqrt{1+\frac{C^2}{r^2}} \tag{17}$$

$$V_w=E_l\left\{\sqrt{R_A^2+C^2}-\sqrt{R_B^2+C^2}-C\ln\left[\frac{R_B(C+\sqrt{R_A^2+C^2})}{R_A(C+\sqrt{R_B^2+C^2})}\right]\right\} \tag{18}$$

where

$$E_l\equiv\sqrt{\frac{-i}{2\pi K\epsilon}} \tag{19}$$

and

$$C^2\equiv\frac{1}{E_l^2}\left(\frac{V_{cr}}{\ln\frac{R_A}{R_B}}\right)^2-R_A^2 \tag{20}$$

Positive corona was used in all the experiments conducted. The corona-starting voltage V_{cr} was measured by noting the wire voltage at which corona current first appeared; with the 6.3-mil wire and 10-inch cylinder used, this value was +7.4 kv. The value of E_l corresponding to a particular wire voltage was then determined by solving equation 18. Since equation 18 is transcendental, a digital computer routine was prepared to aid in data reduction. This routine uses Newton's method for successive approximations to find a value of E_l which fits equation 18. Next, equation 17 was used to calculate the field strength at the particle's location. All measurements were taken with the particles at $r=3.75$ inches.

Table I. Measured Values of Charge/Area Ratio Expressed in Dimensionless Form

Particle Shape	Wire Voltage, Kv				Average Ratio
	20	30	40	48	
Sphere, theoretical	3.00	3.00	3.00	3.00	3.00
Sphere, measured	3.06	3.10	3.16	3.13	3.11
BB shot	3.22	3.26	3.24	3.24	3.24
Sphere-flat	3.14	3.14	3.12	3.00	3.13
Smooth projection	3.24	3.25	3.20	3.14	3.21
Sharp projection	3.33	3.41	3.44	3.44	3.40
Cylinder, axis $\parallel E_0$	3.66	3.60	3.65	3.62	3.63
Cylinder, axis $\perp E_0$	2.87	2.81	2.89	2.89	2.87
Cylinder, hemispherical ends	3.60	3.66	3.61	3.56	3.61

Experimental Results

The measured values of charge were all of the order of 10^{-10} coulombs. To facilitate comparisons among the various shapes, the charge/area ratio was calculated for each particle shape. Seven different particle shapes were investigated:

1. *Spherical particle.* A 3/16-inch diameter ball bearing was used to check the agreement between theoretical and experimental values of charge.

2. *BB shot.* A copper-clad *BB* was used. This particular sample had "diameters" which measured from 0.169 to 0.174 inch, with most of the values being 0.173 inch. The surface area was computed by assuming the *BB* to be a sphere of diameter of 0.173 inch.

3. *Sphere-flat.* A 3/16-inch diameter ball bearing was ground perpendicular to a diameter for a distance of 0.011 inch, creating a flat face on an otherwise spherical particle (this was intended to simulate a characteristic feature of *BB* shot). The surface area of this particle is only 0.33% less than that of the original sphere.

4. *Smooth projection on a sphere.* Half of a 1/16-inch-diameter ball bearing was ground away. The remaining hemispherical segment was then attached to a 3/16-inch diameter sphere with conducting paint (to assure good contact).

5. *Sharp projection on a sphere.* A cylindrical projection 0.036 inch in diameter and pointed at one end was attached to a 3/16-inch sphere. The over-all length of the projection was 0.088 inch, and it added about 10% to the original surface area of the sphere.

6. *Right circular cylinder.* A circular cylinder 3/32 inch in diameter and 1/4 inch long was tested with its long axis first parallel to and later perpendicular to the charging field. These tests provide a check on the orientation dependence of the charging process, which was noted in the analytical work on the charging of spheroids.

7. *Cylinder with hemispherical ends.* The ends of the cylinder described above were rounded off to form hemispheres 3/32 inch in diameter, the over-all length being kept at 1/4 inch. Comparing the charge acquired by this particle with that acquired by the right circular cylinder provides a further test of the effect of surface features of the particle on the charge/area ratio.

The measurements are summarized in Table I, where values are presented of the dimensionless charge/area ratio, obtained by dividing the measured charge/area ratio by the quantity ϵE_0. The measured values for the spherical particle average about 4.0% higher than the value predicted from the spherical-particle theory. The values for the *BB* shot are some 8.0% higher than would be expected from the spherical-particle theory; this value is in substantial agreement with the findings of previous observers. The next three rows of values in the table show that the effect of surface features of the particle on its charge/area ratio are not very great; even a sharply pointed projection produced only about 13% more charge than would be predicted from the spherical-particle theory.

The measurements on the cylinders are of particular interest. For the equivalent orientation, the difference in the values of the charge/area ratios measured for the two cylinders, one with squared ends and the other with hemispherical ends, is less than the experimental error. This is a further indication of the relative insensitivity of the charge/area ratio to the surface features of the particle.

In addition, a comparison between the values measured for the right circular cylinder and values computed for a prolate spheroid of the same approximate shape is remarkable. The right circular cylinder has an "axial ratio" of 2.67, and for the two orientations the measured (dimensionless) charge/area ratios were 3.63 and 2.87 respectively. For a prolate spheroid with an axial ratio of 2.67, the corresponding calculated values are 3.6 and 2.8 respectively.

Conclusions

1. The accumulated evidence points to the conclusion that while the charge

acquired by a particle in a corona discharge can be larger than would be expected from the spherical-particle theory because of the shape of the particle, the increase is unlikely to be of practical significance.

2. The observations of previous investigators on the charging of *BB* shot have been substantiated, but no adequate explanation has been found for the discrepancy between the spherical-particle theory and the experimental results in this case. Either the irregularities of the *BB* increase the surface area by several per cent (which seems rather unlikely), or the charge/area ratio is increased in some way not simulated by any of the other experiments conducted.

3. On the basis of these results it may be concluded that particle shape does not constitute an important design factor in electrostatic precipitation problems for the particles usually encountered.

Nomenclature

a = radius of spherical particle, meters
a = longest semiaxis of ellipsoidal particle, meters
A_{ob} = surface area of oblate spheroid, meters^2
A_{pr} = surface area of prolate spheroid, meters^2
A_s = surface area of sphere, meters^2
b = intermediate semiaxis of ellipsoidal particle, meters
c = shortest semiaxis of ellipsoidal particle, meters
C = quantity defined in equation 20, meters
DVIF = drift velocity improvement factor, dimensionless
$E(r)$ = field strength as a function of r, volts/meter
E_c = electric field strength in the collecting region, volts/meter
E_l = quantity defined in equation 19, volts/meter
E_0 = electric field strength in the charging region, volts/meter
i = corona current per unit length of wire, amperes/meter
K = mobility of ions, meters^2/volt-second
q = charge on a particle, coulombs
q_s = saturation value of particle charge, coulombs
r = radial distance, meters
R_A = radius of corona wire, meters
R_B = radius of outer cylinder, meters
R_ξ = quantity defined in equation 22, meters^3
V = electrostatic potential at a point, volts
V_{cr} = critical voltage for starting corona, volts
V_w = wire voltage with respect to outer cylinder, volts
ϵ = dielectric constant of medium, farads/meter
$\left.\begin{matrix}\xi\\ \eta\\ \zeta\end{matrix}\right\}$ = one of ellipsoidal co-ordinates, meters^2
λ = viscosity of the medium, kilogram/meter-second
σ_1 = surface density of induced charge, coulombs/meter^2
σ_2 = surface density of free charge, coulombs/meter^2

Appendix

When an initially uncharged conducting ellipsoid of principal semiaxes $a > b > c$ is placed in a uniform electric field E_0 with its long axis a parallel to the field, the resulting potential distribution is given in ellipsoidal co-ordinates by[17]

$$V=E_0\sqrt{\frac{(\xi+a^2)(\eta+a^2)(\zeta+a^2)}{(b^2-a^2)(c^2-a^2)}}\left[1-\frac{\displaystyle\int_\xi^\infty \frac{d\xi}{(\xi+a^2)R_\xi}}{\displaystyle\int_0^\infty \frac{d\xi}{(\xi+a^2)R_\xi}}\right] \qquad (21)$$

where

$$R_\xi=\sqrt{(\xi+a^2)(\xi+b^2)(\xi+c^2)} \qquad (22)$$

The distribution of the charge induced on the surface of the ellipsoid is found by calculating the gradient of the potential at the particle surface.

Since the particle surface has the coordinate $\xi=0$ and the normal to the surface is in the ξ direction, the induced surface charge density is[18]

$$\sigma_1=-\epsilon\left[\frac{2R_\xi}{\sqrt{(\xi-\eta)(\xi-\zeta)}}\frac{\partial V}{\partial \xi}\right]_{\xi=0}$$
$$=\frac{-2\epsilon E_0\sqrt{\dfrac{(\eta+a^2)(\zeta+a^2)}{(b^2-a^2)(c^2-a^2)}}}{a\sqrt{\eta\zeta}\displaystyle\int_0^\infty \frac{d\xi}{(\xi+a^2)R_\xi}} \qquad (23)$$

References

1. La Charge des Particules Spheriques dans un Champ Ionise, **M. Pauthenier, M. Moreau-Hanot.** *Journal de Physique et le Radium*, Paris, France, series 7, vol. 3, 1932, pp. 590–613.

2. Measurements of Charge Imparted to Large Particles by Normal and Abnormal Corona, **J. G. Hewitt, Jr.** *Ph.D. Thesis*, Carnegie Institute of Technology, Pittsburgh, Pa., 1956.

3. The Charging of Non-spherical Particles in a Corona Discharge, **P. L. Smith.** *Ph.D. Thesis*, Carnegie Institute of Technology, 1960.

4. Electromagnetic Theory (book), **J. A. Stratton.** McGraw-Hill Book Company, Inc., New York, N. Y., 1941, pp. 207–09.

5. *Ibid.*, pp. 58–59.

6. Handbook of Elliptic Integrals for Engineers and Physicists (book), **P F. Byrd, M. D. Friedman.** Springer-Verlag, Berlin, Germany, 1954, p. 5.

7. Tables of Functions (book), **E. Jahnke, F. Emde.** Dover Publications, Inc., New York, N. Y., fourth edition, 1945, pp. 52–85.

8. A New Method of Expressing Particle Sphericity, **B. C. Aschenbrenner.** *Journal of Sedimentary Petrology*, Tulsa, Okla., vol. 26, 1956, pp. 15–31.

9. Handbook of Engineering Fundamentals (book), **O. W. Eshbach.** John Wiley & Sons, Inc., New York, N. Y., 1936, pp. 2–43.

10. **J. A. Stratton,** *op. cit.*, p. 217.

11. The Physics of Particle Size Measurement. Part I. Fluid Dynamics and the Stokes Diameter, **P. G. W. Hawksley.** *Monthly Bulletin*, British Coal Utilization Research Association, London, England, vol. 15, 1951, pp. 123–25.

12. An Introduction to Electrostatic Precipitation in Theory and Practice (book), **H. E. Rose, A. J. Wood.** Constable & Company, Ltd., London, England, 1956, pp. 57–60.

13. Streamline Flow (book), **H. F. P. Purday.** Constable & Company, Ltd., 1949, pp. 104–06.

14. A New Technique for the Measurement of Corona Field Strength and Current Density in Electrical Precipitation, **P. Cooperman.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, Mar. 1956, pp. 64–67.

15. Some Measurements of Abnormal Corona, **Gaylord W. Penney, Jack G. Hewitt, Jr.** *Ibid.*, vol. 77, July 1958, pp. 319–27.

16. Field-Strength Measurements in Parallel-Plate Precipitators, **J. S. Lagarias.** *Ibid.*, vol. 78, Sept. 1959, pp. 427–33.

17. **J. A. Stratton,** *op. cit.*, pp. 209–11.

18. *Ibid.*, p. 207.

Tunnel Diode D-C Power Converter

H. F. STORM
MEMBER AIEE

D. P. SHATTUCK
ASSOCIATE MEMBER AIEE

Synopsis: The tunnel diode d-c converter serves to generate direct current of a desired potential, when the d-c supply is only a fraction of 1 volt, as may be encountered in connection with silicon solar cells; thermoelectric, thermionic devices; chemical fuel cells; and others. Not only will this paper describe such a converter but also demonstrate how to analyze circuit problems involving a highly nonlinear device, such as a tunnel diode; furthermore how, by intensifying a nonlinearity, the analysis of a nonlinear problem becomes simpler. The price paid for this simplification is reduced accuracy, but this price often appears attractive because of the radical problem-simplification which can be realized.

THE CIRCUIT, which is shown in Fig. 1, consists of a d-c supply voltage E_B, the tunnel diode TD, a transformer with a single primary winding TP, and two equal secondary windings TS_1 and TS_2. The secondary windings together with the ordinary diodes D_1 and D_2 form a full-wave rectifier feeding the load resistor R_L, having a filter capacitor C connected in parallel.

First the converter will be discussed under no-load conditions (secondaries TS_1 and TS_2 open-circuited) and then, by way of superposition, its operation under load.

Rigorous Analysis of Converter—No-Load Conditions

As the secondaries TS_1 and TS_2 of the transformer are considered open-circuited, the active part of the circuit shown in Fig. 2(A) consists of the supply E_B, the tunnel diode TD,[1–9] and the transformer primary TP (between terminals H_1 and H_2), having an inductance L and a resistance R. The resistive characteristic of the tunnel diode is shown in Fig. 2(B); the effects of parallel capacitance and series inductance of the tunnel diode are negligible at the frequency range considered herein. The remaining circuit resistances are designated by r.

With the notations of Fig. 2(A), the circuit voltages are related by

$$E_B - e_T - e_X - i_T(R+r) = 0 \tag{1}$$

When the tunnel diode operates at point A, Fig. 2(B), the exciting voltage e_X follows from equation 1, and is shown in Fig. 2(B) as the horizontal distance (dashed line) from A to the straight line FG; the sign of e_X is positive. With e_X known, the slope of the current follows from equation 2.

$$\frac{di_X}{dt} = \frac{e_X}{L} \tag{2}$$

The slope of the current at A can now be plotted in Fig. 2(C) as the dash-dotted line through A. Although the tunnel diode current i_T and the exciting current i_X are identical in the circuit of Fig. 2, separate symbols are used because a separation of the two currents will become necessary later.

By assuming various, increasing values for i_T, corresponding current slopes are found and, by joining them, the current i_T is obtained as a function of time; see Fig. 2(C), from A to B. Each of the assumed current values yields from Fig. 2(B) a corresponding tunnel diode voltage e_T, exciting voltage e_X, and primary terminal voltage e_E, plotted as functions of time in Fig. 2 (D) and (E).

When reaching point B, Fig. 2(B), the exciting voltage e_X, and hence the current slope, are still positive. Point B, however, represents a maximum and, in its vicinity, current increase is impossible. As a result, the operating point of the tunnel diode shifts to the point C area, where the reactive voltage e_X becomes negative, as does the current slope, equation 2, while the current declines from C to D, Fig. 2(C). The corresponding voltages are shown in Figs. 2 (D) and (E).

When point D is reached, the exciting voltage e_X, Fig. 2(B), and the current slope are still negative. Point D, however, is a minimum and a current decrease in its vicinity is impossible. As a result, the operating point of the tunnel diode shifts to the point A region, Fig. 2(B). The exciting voltage e_X becomes positive again and the current starts to rise, thus completing one cycle of operation.

The period T of this oscillation, which is shown by the distance AD in Fig. 2(C), consists of two subperiods AV and VD. These may or may not be equal, depending on the magnitude of the supply voltage E_D. This analysis, however, is valid, regardless of lengths of the subperiods. Furthermore, the circuit will not oscillate if the straight line FG, Fig. 2(B), intersects the tunnel diode characteristic left of B, or right of D.

Oscillograms taken with the circuit are shown in Fig. 3, and attention is directed to the correlation between theory and practice.

TEST CIRCUIT DATA

Transformer Windings

Primary turns $N_P = 10$; primary resistance $R = 0.166$ ohm; primary inductance $L = 58$ μh (microhenrys); secondary turns (between neutral and end) $N_S = 213$; secondary resistance $R_S = 0.88$ ohm.

Transformer Core

Material Ferramic *H;* outside diameter 3/4 inch; inside diameter 1/2 inch; height 1/4 inch; length = 5 cm (centimeters); area = 0.2 cm².

Tunnel Diode

Ge (Germanium), see characteristic in Fig. 2(B); $I_{T,\max} = 0.118$ amp (ampere); $I_{T,\min} = 0.015$ amp; parallel capacitance 150 μμf (micromicrofarads); series inductance 6 nh (nanohenrys).

$C = 2$ μf; $E_B = 0.235$ volt; $r = 0.139$ ohm.

Obtaining Oscillation Period T—No-Load Conditions

Where the period T consists of two equal subperiods and where it needs to be known only approximately the analysis can be greatly simplified.

Curved branches OB and DC of the tunnel diode characteristic are replaced

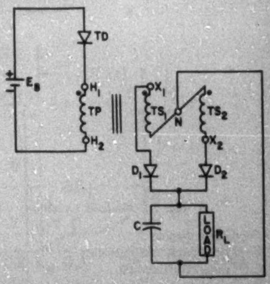


Fig. 1. Tunnel diode converter circuit

Paper 61-149, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 4, 1960; made available for printing December 1, 1960.

H. F. STORM and D. P. SHATTUCK are with the General Electric Company, Schenectady, N. Y.

by two straight and vertical branches through Q and S, Fig. 4, and positioned to make the areas $AA'Q = BB'Q$, and $DD'S = CC'S$; this positioning can be quickly accomplished by a transparent ruler and visual inspection. Thus, the current-dependent tunnel diode voltages are replaced by the constant tunnel diode voltages $E_{T,1}$ and $E_{T,2}$. Similarly, the line FG, Figs. 2(B) and 4, is replaced by the vertical line through W to make the triangles WXX' and WYY' equal. Thus, variable voltage $E_B - i_T(R+r)$ is replaced by constant voltage E_B'. With these simplifications, the exciting voltage e_X becomes constant and equals $E_B' - E_{T,1}$ for increasing currents, and $E_B' - E_{T,2}$ for decreasing currents. The current derivatives are now

$$\frac{di_X}{dt} = \frac{E_B' - E_{T,1}}{L} \tag{3A}$$

for increasing current, and

$$\frac{di_X}{dt} = \frac{E_B' - E_{T,2}}{L} \tag{3B}$$

for decreasing current.

Subperiod T_1 for increasing the current from $I_{T,\min}$ to $I_{T,\max}$ follows from an integration of equation 3(A)

$$T_1 = \frac{L}{E_B' - E_{T,1}} \, (I_{T,\max} - I_{T,\min}) \tag{4}$$

For equal subperiods for increasing and decreasing currents

$$E_B' - E_{T,1} = E_{T,2} - E_B' \tag{5}$$

resulting in

$$E_B' = \frac{1}{2}\,(E_{T,1} + E_{T,2}) \tag{6}$$

Designating

$$\Delta I_T = I_{T,\max} - I_{T,\min} \tag{7A}$$

$$\Delta E_T = E_{T,2} - E_{T,1} \tag{7B}$$

produces from equations 4, 6, 7(A), and 7(B), the period T, $(T = 2T_1)$ for one full oscillation.

$$\boxed{T = 4L\,\frac{\Delta I_T}{\Delta E_T}} \tag{8}$$

As quantities ΔI_T and ΔE_T are readily.

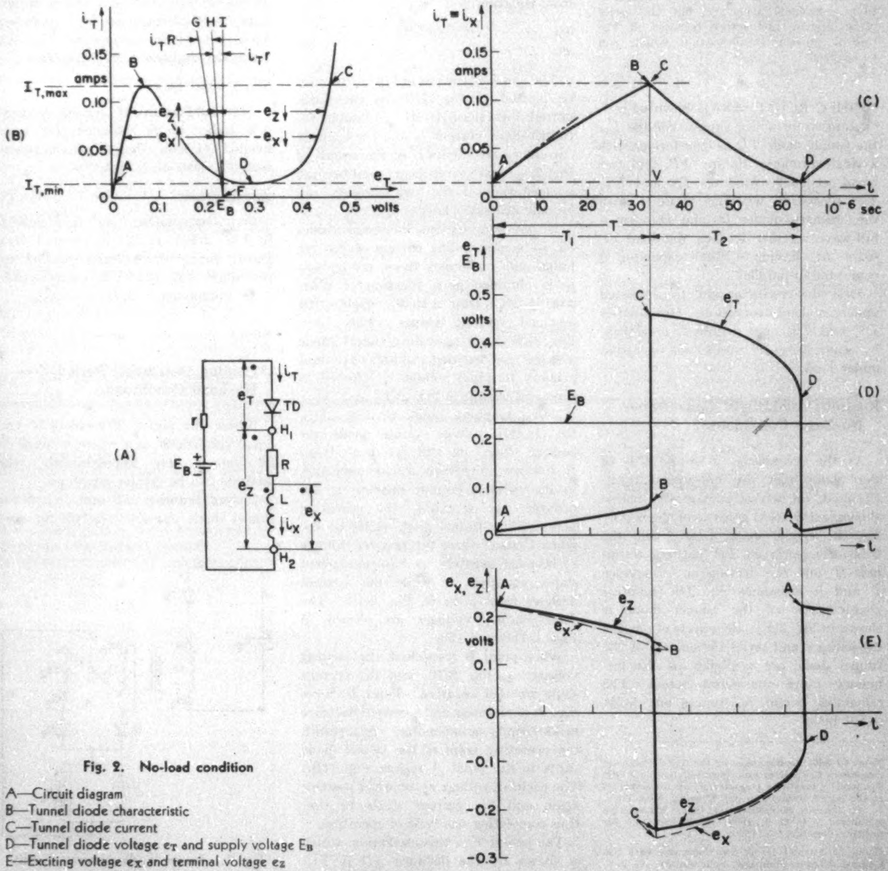

(A)
r
E_B
e_T
i_T
TD
H_1
R
e_Z
L
i_X
e_X
H_2
(B)
i_T
0.15
amps
0.10
0.05
0
$I_{T,max}$
$I_{T,min}$
B
A
D
F
C
G H I
$i_T R$
$i_T r$
e_Z
e_X
$-e_Z$
$-e_X$
e_T
0.1
0.2
E_B
0.3
0.4
0.5
volts
(C)
$i_T = i_X$
0.15
amps
0.10
0.05
0
A
B
C
D
V
t
10
20
30
40
50
60
10^{-6} sec
T
T_1
T_2
(D)
e_T
E_B
0.5
volts
0.4
0.3
0.2
0.1
0
C
e_T
D
E_B
B
A
A
t
(E)
e_X, e_Z
0.2
volts
0.1
0
-0.1
-0.2
-0.3
A
e_Z
e_X
B
A
D
t
C
e_Z
e_X

Fig. 2. No-load condition

A—Circuit diagram
B—Tunnel diode characteristic
C—Tunnel diode current
D—Tunnel diode voltage e_T and supply voltage E_B
E—Exciting voltage e_X and terminal voltage e_Z

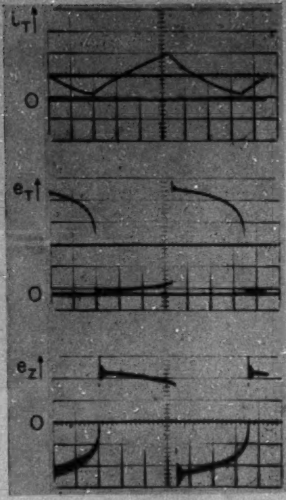


Fig. 3. Test data taken under same conditions as theoretical development of Fig. 2

i_T—0.06 amp/division
e_T, e_Z—0.1 volt/division
t—10 μsec/division

available from the tunnel diode characteristic, the period of oscillation can be calculated quickly.

In the present case, $\Delta I_T = 0.103$ amp, $\Delta E_T = 0.404$ volt, $L = 58$ μh, resulting in a calculated period $T = 59.3$ μsec (microseconds); the measured period is 65.8 μsec.

Analysis of Converter—with Load

As seen from the previous analysis and especially from Fig. 2(C), the exciting current i_X has an alternating component; hence, the transformer core carries an alternating flux component. Returning to the complete circuit shown previously in Fig. 1 and in more detail in Fig. 5(A), voltage E_L, determined by the turns-ratio N_S/N_P, is seen to develop across the load resistor R_L.

The ensuing circuit analysis is based on assumptions: The forward resistance of the diodes D_1, D_2 is negligible; reverse resistance is infinite. The time constant of CR_L is long enough to make the voltage ripple on the load resistance R_L negligible. The turns-ratio N_S/N_P between each of the secondary windings with respect to the primary winding is unity, and the coupling among these windings is also unity. The effect of secondary winding resistance is accounted for by referring it to the primary of the transformer.

Suppose a load voltage E_L exists on the capacitor-resistor combination CR_L, Fig. 5(A), and a secondary current $i_{S,1}$ is flowing. Because of the prior assumptions, load voltage E_L equals the exciting voltage e_X

$$e_X = E_L \tag{9}$$

Replacing e_X in equation 1 with equation 9 solves for the tunnel diode voltage e_T. This is graphically accomplished in Fig. 5(B) by drawing line JK parallel to FG, and is horizontally displaced by load voltage E_L; intersection K of line JK with the tunnel diode characteristic determines the tunnel diode voltage e_T and current i_T. With the polarities of Fig. 5(A), secondary current $i_{S,1}$ is the difference between the primary (tunnel diode) current i_T and the exciting current i_X

$$i_{S,1} = i_T - i_X \tag{10}$$

In obedience to equations 2 and 9, the exciting current i_X is

$$i_X = \frac{1}{L}\int e_X dt = \frac{E_L}{L}\,t + \text{constant} \tag{11}$$

At $t=0$, $i_X = I_{T,\min}$ and from equation 11

$$i_X = \frac{E_L}{L}\,t + I_{T,\min} \tag{12}$$

See Fig. 5(C_1). And, in connection with equation 12, the secondary current $i_{S,1}$ becomes

$$i_{S,1} = i_T - \left(\frac{E_L}{L}\,t + I_{T,\min}\right) \tag{13}$$

See Fig. 5(C_2). This mode persists until the exciting current i_X equals the tunnel diode current i_T, at which point the secondary current $i_{S,1}$ becomes zero. Denoting this instant of time t_1, equation 13 yields

$$t_1 = \frac{L}{E_L}\,(I_{T,K} - I_{T,\min}) \tag{14}$$

where $I_{T,K}$ signifies the tunnel diode current at K of the tunnel diode characteristic Fig. 5(B); during $0<t<t_1$, the operating point remains at K.

For $t_1<t<T_1$, the excitation proceeds precisely as first explained, the exciting voltage e_X being the horizontal distance between the tunnel diode characteristic KB and the straight line FG. The concommitant voltages e_T, e_X, $e_{S,1}$, $e_{S,2}$, and e_Z are derived from Fig. 5, (B), (C_1), and (C_2), and are shown in Fig. 5, (D) and (E).

When reaching the maximum B, Fig. 5(B), the exciting voltage e_X is still positive; hence, the exciting current still has a tendency to increase. As a result, the operating of the tunnel diode shifts from B to the CLD branch of the tunnel diode characteristic. This shift causes a reversal in the sign of e_X, and the core flux begins to drop. Now, a secondary voltage $-e_{S,2}$ is produced, sending a current $i_{S,2}$ into the CR_L circuit. In view of the assumptions

$$-e_{S,2} = E_L = -e_X \tag{15}$$

Substituting E_L for $-e_X$ in equation 1 solves for e_T. This is accomplished graphically by intersecting straight line MN with the tunnel diode characteristic Fig. 5(B); the line MN is parallel to FG, and displaced horizontally to the right of FG by the amount of E_L. The tunnel diode voltage and current are indicated by point L of Fig. 5(B).

Analogous to equation 11, the exciting current i_X becomes

$$i_X = -\frac{E_L}{L}\,(t - T_1) + \text{constant} \tag{16}$$

At $t=T_1$, $i_X = I_{T,\max}$, equation 16 yields

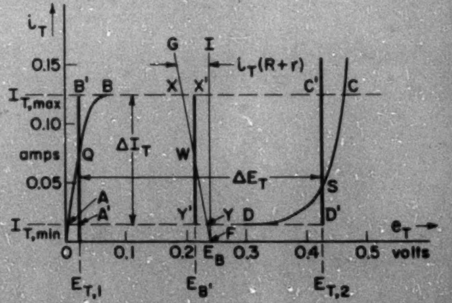


Fig. 4. Simplification of Fig. 2(B) for rapid determination of oscillation period, equation 8

$$i_X = -\frac{E_L}{L}(t-T_1)+I_{T,\max} \qquad (17)$$

With the diode D_2 conducting, and the polarities shown in Fig. 5(A)

$$i_{S,2} = -i_T + i_X \qquad (18)$$

From equations 17 and 18

$$i_{S,2} = -i_T - \frac{E_L}{L}(t-T_1)+I_{T,\max} \qquad (19)$$

Tunnel diode current i_T, exciting current i_X, and secondary current $i_{S,2}$ are shown in Fig. 5, (C_1) and (C_2).

This mode continues until $t=t_2$ where the exciting current i_X equals the tunnel diode current i_T, and the secondary current $i_{S,2}$ vanishes. Letting $i_{S,2}=0$, t_2 is found from equation 19

$$t_2 = \frac{L}{E_L}(I_{T,\max}-I_{T,L})+T_1 \qquad (20)$$

where $I_{T,L}$ is the current at point L of Fig. 5(B).

For $t_2<t<T$, the excitation proceeds again as described in the first section, the exciting voltage e_X being the horizontal difference between the tunnel diode characteristic DL and the line FG. When the minimum point D is reached, the operation transfers to branch AKB of the tunnel diode characteristic, and a new cycle begins. Thus, all voltages, currents, and subperiods T_1, T_2 are accounted for. The waveshapes in Fig. 5 are drawn for $E_B=0.234$ volt, $R_L=2$ kilohms.

Oscillograms of the loaded oscillator are shown in Fig. 6, for $E_B=0.25$ volt, $R_L=10$ kilohms, $f=14.3$ kc.

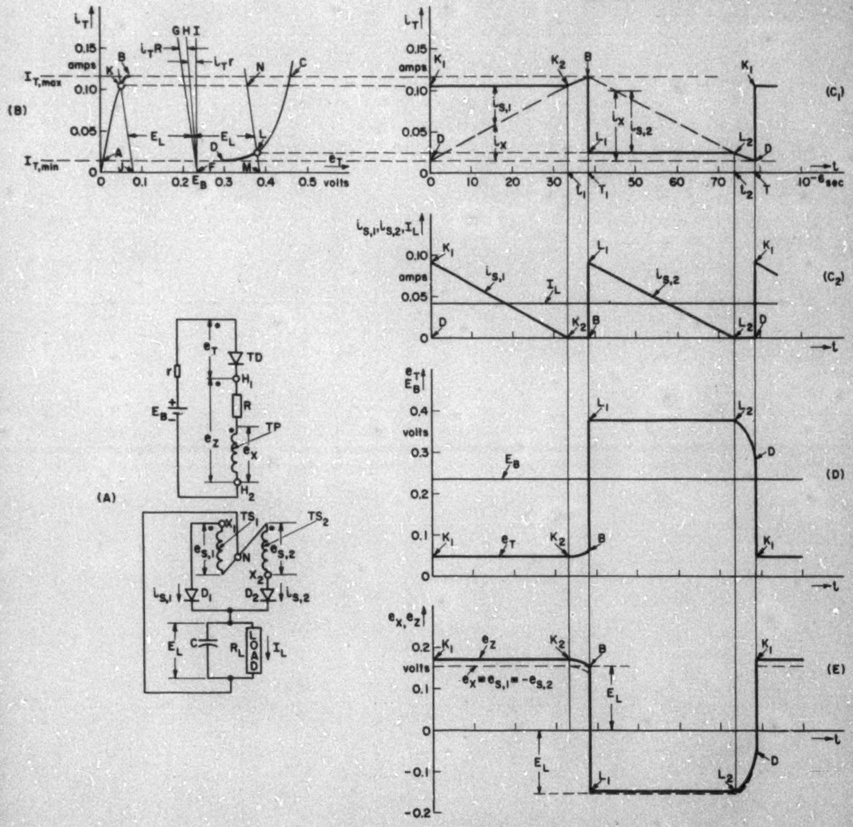

(B)
i_T
0.15
amps
0.10
0.05
0
$I_{T,\max}$
$I_{T,\min}$
K
B
A
J
G H I
$i_T R$
$i_T r$
N
C
E_L
D
L
F
M
e_T
0 0.1 0.2 E_B 0.3 0.4 0.5 volts
(C_1)
K_1 K_2 B K_1
$i_{S,1}$
i_X
$i_{S,2}$
L_1 L_2 D
0 10 20 30 t_1 T_1 50 60 70 t_2 T 10^{-6} sec
t
(C_2)
$i_{S,1}, i_{S,2}, I_L$
0.10
amps
0.05
I_L
K_2 B L_2 D
(D)
e_T
E_B
0.4
volts
0.3
0.2
0.1
L_1 L_2
e_T
(E)
e_X, e_Z
0.2
volts
0.1
-0.1
-0.2
e_Z
$e_X = e_{S,1} = -e_{S,2}$
E_L
(A)
e_T
TD
r
H_1
R
E_B
TP
e_Z
e_X
H_2
TS_1
TS_2
X_1
N
X_2
$e_{S,1}$
$e_{S,2}$
$i_{S,1}$
D_1
D_2
$i_{S,2}$
E_L
C
R_L
LOAD
I_L

Fig. 5. Converter with load ($R_L = 2k$)

A—Circuit diagram
B—Tunnel diode characteristic
C_1—Tunnel diode current i_T
C_2—Transformer secondary currents $i_{S,1}$, $i_{S,2}$, average load current I_L
D—Tunnel diode voltage e_T, supply voltage E_B
E—Exciting voltage e_X, terminal voltage e_Z, load voltage E_L (d-c)

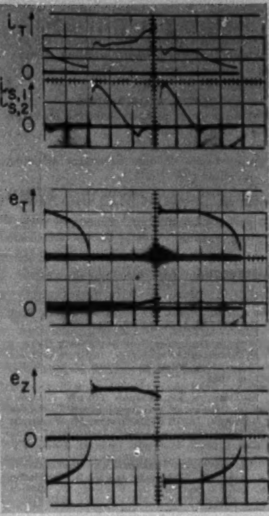


Fig. 6. Oscillograms of loaded converter

$E_B=0.25$ volt, $R_L=10$ kilohms, $f=14.3$ kc
i_T—0.06 amp/division
$i_{S,1}$, $i_{S,2}$—0.001 amp/division
e_T, e_X—0.1 volt/division
t—10 μsec/division

Operation Near Maximum Output—Threshold Performance

Fig. 5, (B), (C_1), and (C_2), indicates that the transformer secondary currents and hence the load current I_L increase if the load voltage E_L is reduced. The largest load current is obtained when the supply voltage E_B and the load voltage E_L are reduced until the tunnel diode operates for one half-period at point B and the other half-period at point D of Fig. 5(B). This type of operation is called threshold operation because the slightest reduction of E_B or E_L will cause improper operation of the converter.

The converter power output P_L is the product of E_L and I_L:

$$P_L=E_LI_L \tag{21}$$

where

$$I_L=\frac{1}{T}\left[\int_0^{t_1} i_{S,1}dt+\int_{t_1}^{t_2} i_{S,2}dt\right] \tag{22}$$

If operating points of the tunnel diode are moved from A and C toward B and D, respectively, Fig. 5(B), the power output P_L increases until B and D are reached, or slightly before. The converter will be analyzed for threshold operation, with the assumption of winding and lead resistances $R+r$ being negligible. This simplifies the analysis, and yet produces useful data for approximating converter operation at maximum power output.

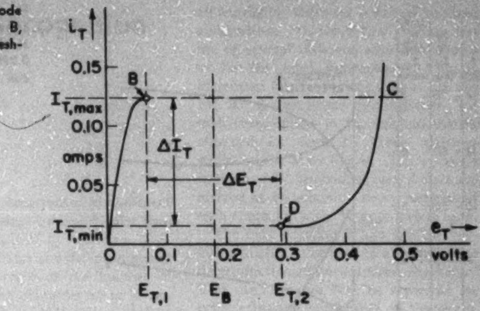


Fig. 7. Tunnel diode operating points B, D at stability threshold

The voltage ΔE_T is obtained from equation 7B, with $E_{T,1}$ and $E_{T,2}$ determined from the extreme points B and D of the tunnel diode characteristic, Fig. 7. From inspection of Fig. 5(B), the load voltage E_L becomes simply $1/2\Delta E_T$. A turns-ratio N_S/N_P different from unity gives

$$\boxed{E_L=\frac{1}{2}\Delta E_T\frac{N_S}{N_P}} \tag{23}$$

Load current I_L follows from equations 7A and 22

$$\boxed{I_L=\frac{1}{2}\Delta I_T\frac{N_P}{N_S}} \tag{24}$$

Load resistance R_L follows from equations 23 and 24

$$\boxed{R_L=\frac{E_L}{I_L}=\frac{\Delta E_T}{\Delta I_T}\left(\frac{N_S}{N_P}\right)^2} \tag{25}$$

and frequency is derived from equation 8

$$\boxed{f=\frac{1}{T}=\frac{1}{4L}\frac{\Delta E_T}{\Delta I_T}} \tag{26}$$

Finally, power output P_L into the load is obtained from equations 23 and 24

$$\boxed{P_L=E_LI_L=\frac{1}{4}\Delta E_T\cdot\Delta I_T} \tag{27}$$

Comparison of calculated and measured values is exhibited in Table I.

Circuit Characteristics

Circuit characteristics are shown by varying load resistance R_L. This is given in Fig. 8.

Total efficiency of this converter near threshold operation is 33%. Loss in the tunnel diode is calculated at a maximum of 48% input power; remaining losses are 19%.

Loss in the tunnel diode can be lowered by reducing ratios $E_{T,1}/\Delta E_T$ and $I_{T,\text{min}}/\Delta I_T$ of Fig. 7. Recently developed GaAs (gallium arsenide) diodes[10] have a current ratio ($I_{T,\text{min}}/\Delta I_T$) of only one-third the ratio of the Ge diode in Fig. 2(B), leading to higher converter efficiencies. If these two ratios approach zero, the tunnel diode efficiency becomes 100%. This statement is of interest when comparing the converter to a tunnel diode oscillator operating within the negative-resistance region, BD of Fig. 2(B); such an oscillator has a maximum theoretical efficiency of only 50%.

Tunnel diodes are available with $I_{T,\text{max}}$ current ranges from milliamperes to several amperes. The voltage swing

Table I. Performance of Converter at Maximum Power Output

Quantity	Equation	Calculated	Measured
E_L	23	$\frac{1}{2}0.225.21.3=2.4$ volts	2.8 volts
R_L	25	$\frac{0.225}{0.103}.21.3^2=1$ kilohm	1.6 kilohms
P_L	27	$\frac{1}{4}$ 0.225.0.103=5.8 megawatts	5 megawatts
f	26	$\frac{10^6}{4.58}\frac{0.225}{0.103}=9.4$ kc	10.9 kc

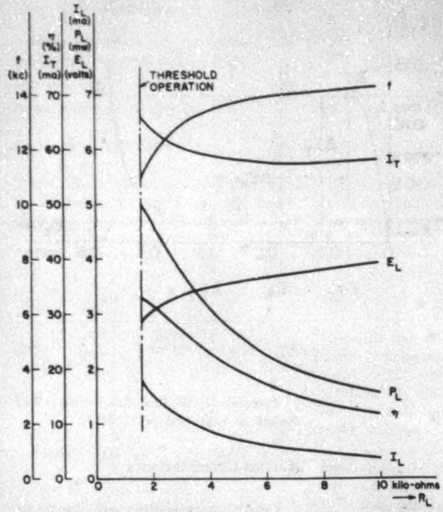


Fig. 8. Circuit characteristics, Fig. 5(A), with E_B = 0.228 volt, as function of load resistance R_L

depends on the type of semiconducting material; a silicon tunnel diode may have a voltage swing of approximately 50% more than the Ge tunnel diode of Fig. 2(A), and a GaAs tunnel diode will double the voltage of the Ge diode.

Conclusions

The described power converter is capable of operating from a d-c supply of only a fraction of 1 volt. The circuit components are simple and rugged. Despite severe circuit nonlinearity, simple equations for performance prediction were derived.

Nomenclature

E_B = supply voltage, volts, average
E_L = load voltage, volts, average
ΔE = average, as marked
$e_{S,1}$, $e_{S,2}$ = transformer secondary voltages, volts, instantaneous
e_T = tunnel diode voltage, volts, instantaneous
e_X = exciting voltage, volts, instantaneous
e_Z = primary terminal voltage $(e_X + i_T R)$, volts, instantaneous
f = frequency, cycles per second
I_L = load current, amp, average
$I_{T,K}$ = tunnel diode current at point K, amp
$I_{T,L}$ = tunnel diode current at point L, amp
ΔI = average, as marked
$i_{S,1}$, $i_{S,2}$ = transformer secondary current, amp, instantaneous
i_T = tunnel diode current, amp, instantaneous
i_X = exciting current, amp, instantaneous
L = inductance of primary winding, henrys
N_P = primary turns
N_S = secondary turns
P_L = load power, watts, average
R = resistance of primary winding on transformer, ohms
R_L = load resistance, ohms
R_S = resistance of one secondary winding, ohms
r = lead and source resistance of primary circuit, ohms
T = period of oscillation, seconds
t = time, seconds

References

1. NEW PHENOMENON IN NARROW GE *p-n* JUNCTIONS, **L. Esaki.** *Physical Review*, New York, N. Y., vol. 109, 1958, pp. 603–04.

2. TUNNEL DIODES AS HIGH FREQUENCY DEVICES, **H. S. Sommers.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 47, no. 7, July 1959, p. 1201.

3. GERMANIUM AND SILICON TUNNEL DIODES—DESIGN, OPERATION AND APPLICATION, **I. A. Lesk, N. Holonyak, Jr., U. S. Davidsohn, M. W. Aarons.** *WESCON Convention Record*, Institute of Radio Engineers, pt. 3, ECG-441, 1959, pp. 9–31.

4. TUNNEL DIODE—CIRCUITS AND APPLICATIONS **I. A. Lesk, N. Holonyak, Jr., U. S. Davidsohn.** *Electronics*, New York, N. Y., Nov. 27, 1959, p. 60.

5. TUNNEL DIODE OPERATION AND APPLICATION, **I. A. Lesk, J. J. Suran.** *Electrical Engineering*, vol. 79, no. 4, Apr. 1960, pp. 270–77.

6. TUNNEL DIODES AND THEIR USE AS MULTIFUNCTION CIRCUIT ELEMENTS, **J. J. Tieman.** *Digest of Technical Papers*, International Solid State Circuits Conference, L. Winner, New York, N. Y., 1960, pp. 8–9.

7. HIGH FREQUENCY NEGATIVE RESISTANCE CIRCUIT PRINCIPLES FOR ESAKI-DIODE APPLICATIONS, **M. E. Hines, W. W. Anderson.** *Ibid.*, pp. 12–13.

8. A NEW DEVICE USING THE TUNNELING PROCESS IN NARROW *P-N* JUNCTIONS, **L. Esaki, Y. Miyahara.** *Solid-State Electronics*, New York, N. Y., vol. 1, no. 1, Mar. 1960, pp. 13–21.

9. PHYSICAL PRINCIPLES OF THE ESAKI DIODE AND SOME OF ITS PROPERTIES AS A CIRCUIT ELEMENT, **R. A. Pucel.** *Ibid.*, pp. 22–33.

10. GALLIUM ARSENIDE TUNNEL DIODES, **N. Holonyak, Jr., I. A. Lesk.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., Aug. 1960, pp. 1405–09.

Errata

"Symmetric Switching Functions—Matrix Logic V" by E. J. Schubert, published in AIEE Transactions, part I (Communication and Electronics), volume 78, 1959 (January 1960 section), pages 1085 and 1086.

On page 1085 the second paragraph should read "In Table III..."

In the fourth paragraph the last line should read "...of lower degree" instead of "higher."

In equation 11 the last term should read $k \leq m$ instead of $k \geq m$.

In the sixth paragraph the last line should read "...rules 1 and 2."

On page 1086 the last line of the first paragraph should read "...any more, except if either

$$k_1 = p,\ k_2 = \mathcal{L} \qquad (14D)$$

or

$$k_1 = k_2 = 1 \qquad (14F)$$

In making these changes the author wishes to acknowledge the assistance of John Earle of the International Business Machines Corporation.

Rotating Raster Character Recognition System

R. W. WEEKS
NONMEMBER AIEE

CONSIDERABLE EFFORT in the past few years has been spent in the theoretical development of statistical decision theory.[1] In applying this theory to pattern recognition two major problems become immediately apparent. The first problem is to decide which statistical properties are significant (and reasonably independent) for the patterns involved. The second is to choose the means of instrumentation in using these statistics for decision making.

Briefly, this paper describes the procedure used in applying this theory to a rotating raster character recognition system. The first section generally describes the process, the second presents the results, the third describes a specific system designed to instrument the process. N. M. Abramson has presented a more complete and mathematically rigorous treatment of the decision making process.[2]

Statistical Analysis as Applied to the Rotating Raster Character Recognition System

As has been indicated, first a number of statistical properties of the character to be recognized are studied and tested. An example for the technique presented here is a determination of the number of crossings encountered when the light beam moves across the character in a particular direction and position, as shown by the dashed line in Fig. 1. In this case there are two crossings; however, there might be a different number of crossings if the 8 had been written differently or if some other character had been used. The probability of occurrence of different numbers of crossings by a particular scan line can be measured by testing a large number of characters and determining the fractional occurrence of the number of crossings for each character to be recognized. A matrix for this statistic might be as shown in Fig. 2. The number in the matrix element is the probability of the occurrence of the number of crossings encountered, as determined by the column, when scanning the character represented by that row. This probability is arrived at by scanning a very large number of characters. In the terminology of probability theory, this is the conditional probability of occurrence of statistic k given the occurrence of character C. That is, $P(k/C)$. Once this matrix has been established it can be used to determine the probable identity of an unknown character input to the scanner. The unknown character is scanned and a particular number, say k, of crossings is encountered. Merely by inspecting the column, k, and selecting the element which has the largest conditional probability, the character most likely to give this particular number of crossings for this scan is determined. On a quantitative basis the actual probability that the scanned character was any particular character can be calculated using Bayes theorem of conditional probability. This theorem states that the probability of occurrence of a particular character C, if statistic k has occurred, is the conditional probability $P(k/C)$, described previously, divided by the probability of occurrence of statistic k for the system. This last probability, $P(k)$, can be calculated easily, since it is merely the sum of all of the conditional probabilities in column k. Mathematically, this is:

$$P(C/k)=\frac{P(k/C)}{P(k)}=\frac{P(k/C)}{\sum_{c=0}^{9} P(k/C)}$$

The above arguments, strictly speaking, hold only for the case where all characters occur with equal likelihood, which is the case for the ten digits as nearly as can be determined, except perhaps for some special cases. Obviously, a single scan line is inadequate to distinguish between all ten digits; there are bound to be some ambiguities. To solve this problem the use of several different scan lines seems to be the most appropriate. This is illustrated by the various dashed lines in Fig. 3. A matrix can be constructed for each scan similar to the original matrix obtained for scan 1. These matrices might appear as on Fig. 4. The numbers in the matrix elements of each matrix are arrived at by scanning many samples of each character with the appropriate scan.

The next question is how are the various statistics to be used in a quantitative determination of an unknown character? This question can best be answered by considering the probability of the simultaneous occurrence of N separate things. This is merely the product of the probability of occurrence of event 1, times the probability of occurrence of event 2, times the probability of occurrence of event 3, etc., times the probability of occurrence of event N. This is rigorously true in a mathematical sense only if the occurrence of the various events are independent of one another. This property is assumed to apply to the various scan lines.

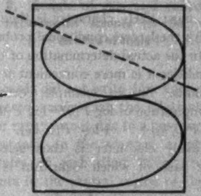


Fig. 1. Typical scan line across an 8. (The rectangle surrounding the 8 is visible to the scanner)

	NUMBER OF CROSSINGS			
CHARACTER SCANNED	0	1	2	3
0	0	.01	.98	.01
1	.2	.7	.1	0
2				
3		etc.		
4				
5				
6				
7				
8				
9	0	.2	.5	.2

Fig. 2. Typical matrix for the statistics of the crossings encountered on one scan line, for the case of the ten digits $k=0, 1, 2, 3$, $C=0, 1, \ldots, 9$

Paper 60-1005, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, San Diego, Calif., August 8–12, 1960, and represented for discussion only at the Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted May 26, 1960; made available for printing November 10, 1960.

R. W. WEEKS is with the Montana State College, Bozeman, Mont.

This paper is the result of the efforts of many people at the San Jose Laboratory over a 2-year period. The early theoretical work was done by E. A. Quade and P. Baxendale, experimental work by M. V. Mollman, and the programing for the IBM *650* computer was done by A. C. Swain. The successful completion of the project is attributable in large part to the advice, on both theory and experiment, by E. A. Quade, W. E. Abramson, and many others.

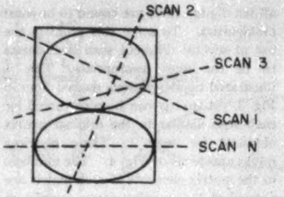


Fig. 3 (left). Typical scan lines 1, 2, 3, ..., N

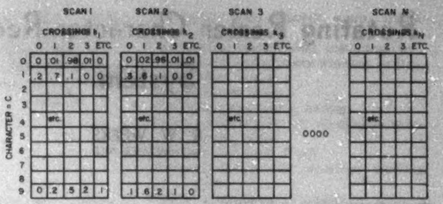


Fig. 4 (right). Typical matrices of N scan lines

Assume some unknown character has been scanned and on each scan line some number of crossings was encountered. For scan 1, ten conditional probabilities can be computed, one for each character. This same computation can be performed for scan 2, scan 3, ..., scan N. The computations for scan 2, for example, would be: $P(0/k_2)$, $P(1/k_2)$, $P(2/k_2)$, ..., $P(9/k_2)$. From these N sets of conditional probabilities the conditional probability that the unknown character is any particular character can be calculated. These conditional probabilities are:

$$\text{Cond } P(0)=P(0/k_1)\times P(0/k_2)\times P(0/k_3)\times \ldots \times P(0/k_N)$$

$$\vdots \qquad \vdots$$

$$\text{Cond } P(9)=P(9/k_1)\times P(9/k_2)\times P(9/k_3)\times \ldots \times P(9/k_N)$$

The most likely character is then the one with the largest conditional probability. In the actual determination of these probabilities it is more convenient to use logarithms. As an example, consider the computation of $\log P(0)=\log P(0/k_1)+\log P(0/k_2)+\ldots+\log P(0/k_N)$. The most likely character is then selected on the basis of which logarithm is the largest.

In the system used in the experiment performed at the San Jose Laboratory a total of 36 individual scan lines was used, or $N=36$. These were selected in such a manner as to simplify the electronics. The digits were scanned in television fashion with a raster consisting of six lines uniformly distributed over the digit at six different angles uniformly spaced 30 degrees apart. The six rasters are shown in Fig. 5. This differs slightly from the system described in section 3 of this paper in which eight rasters of four lines each are used. The system described in section 3 is a redesigned version of the one actually built and differs from it in two major respects. In the actual system, the information on number of crossings was stored in a punched card rather than in a core storage unit, and the decision process was simulated on an IBM *650* computer.

The analysis just referred to is described in terms of a particular system and its associated statistics. Independent statistics obtained in any other fashion would be handled in the same way. A second experiment using the crossings-counting technique was performed. The statistics used were derived from the order in which the types of crossings occurred rather than from the number of crossings in a particular scan line. In other words, the character was scanned a large number of times at each angle and only the sequence of types of crossings encountered was recorded. For example, the following crossings might be encountered at a particular angle of scan:

0000 11111 2222 3333 2222 3333 11111 000
(0) (1) (2) (3) (2) (3) (1) (0)

All that was recorded for statistical use was the sequence: 01232310.

The statistical matrices derived from this information were based on the kind of crossings other than zero encountered first, the kind of crossings encountered second, and the kind of crossings encountered third. This information was gathered for 12 angles giving a total of 12×3 or 36 different statistics. These statistics, while appearing to eliminate completely the size and registration problems, proved not to be useful for recognition, primarily because the three statistics obtained for any particular angle are far from being independent. This is obvious since the probability of changing from one type of crossing to another which differs by plus or minus one crossing per scan is very high, while the probability of changing to a type of crossing differing by two or more crossings per scan is very small. This indicates the dependency of the various statistics used in this second experiment.

Discussion of Experimental Technique and Results

The sample characters used in collecting the statistical data were printed on special cards with ten rectangles 3/16 inch×1/4 inch printed in light blue ink, Fig. 6. The light blue ink was of such a shade as to be indistinguishable from the *J*-Natural card stock by the flying-spot cathode-ray tube scanner. The crt (cathode-ray tube) used was a DuMont *5AQP15*. Two kinds of sample characters were collected, hand-printed and machine-printed. The hand-printed cards were obtained by having a group of 100 students from a local college fill out 20 cards each. The students were told to use the style of characters of which they were given a

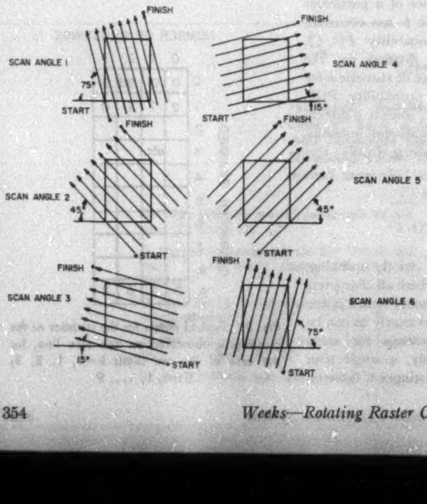


Fig. 5. Rasters used at each of the six scanning angles. Video amplifier was blanked out during the retrace period

sample and to try to make each digit fill out the small rectangular box. No further restraints were imposed. The students used IBM mark-sensing pencils which gave adequate signal-to-noise ratio in the photoreflective scanner. The machine-printed cards were printed in black ink by a standard printing process. The only criterion for selecting the style of characters was that the height-to-width ratio be such that the digits would more or less fill out the rectangular box, Fig. 7.

The photoreflective scanner was mounted on an *026* card punch so that card feeding could be made automatic. The sample cards were fed into the scanner and the crossing information for each digit was punched into six data cards on another *026* punch, one card containing the information for each of the six scanning angles. A sample card is shown in Fig. 8. The raster used for scanning consisted of 64 individual lines. The number of crossings encountered on each line, varying from zero to three, was punched into the appropriate column on the data card. The raster was adjusted so that it overlapped the character being scanned by about 25% at both the top and bottom. This was done to allow for any misregistration. Only six out of the 64 scan lines were used in the recognition process. These were selected in the following manner by the computer.

The computer was programed to scan each card and locate the first crossings count greater than 0, which, because of the high resolution of the scanner, was usually 1. The computer then located the last crossings count other than 0 on the card, and divided the intervening columns into seven equal groups. The last columns of groups 1, 2, 3, 4, 5, and 6 were used as the six scan lines equally spaced across the character. The six rasters, one for each angle, shown in Fig. 5, thus produced 36 scan lines.

In collecting the statistical data for the matrices used in the recognition process, a total of 36 matrices were required, one for each scan line. Each matrix consisted of 40 elements, one element for each digit, 0–9, for each type of crossing, a 0, a 1, a 2, or a 3. This requires a total of 36×40, or 1,440 individual statistics. These were collected by scanning a large number of samples of each digit and counting in the appropriate register of the computer the number of times a particular number of crossings occurred for each of the 36 scan lines. When this process was completed the totals in the 1,440 registers ranged from 0 to the number of samples of each digit scanned. Three hundred hand-printed and 100 machine-printed samples were scanned. The number of digits was then normalized so that the count in each register was based on 1,000 samples. A single count was then added into each register so that all registers had one or more counts. These counts were then converted to logarithms, ranging from zero to three, for use in the decision process. To test the recognition ability of this technique, additional samples of 1,000 handwritten digits and 1,000 machine-printed digits were obtained to test against the previously obtained statistics. These samples were scanned in the same manner as the original samples used to collect the statistics, and six data cards were punched for each digit.

The computer then selected the 36 individual scan lines for each test digit exactly as was done on each statistical sample digit. The crossings count in each of the 36 columns determined which statistical matrix register was involved for each of the ten digits. In other words, for each scan line, ten matrix registers were selected, one for the 0, one for the 1, . . ., one for the 9. The logarithmic quantity stored in the matrix selected for the 0 was added into an accumulator for the 0, that selected for the 1 was added into an accumulator for the 1, etc. This process was repeated for each of the 36 scan lines. At the conclusion of this process the ten accumulators contained quantities that were proportional to the logarithms of the probabilities that the tested digit was a 0, 1, 2, . . ., 9. The accumulator of the digit that was being tested was then compared with the accumulator which had the highest value of the remaining nine digits. If the correct accumulator had a value greater than the highest of the other nine, a correct recognition had been made, if not, a misrecognition had occurred. The difference between the two accumulator values was used to establish a rejection level. In other words, if the difference was below some predetermined value, the digit was rejected rather than recognized or misrecognized. This general technique for making a decision is based on the principles of statistical decision theory, and as such is useful for any set of 36 statistics that can be considered as independent. The results obtained by this technique were as described in the Appendix (Tests 1 and 2).

The scanner appeared to have a somewhat poorer video signal for the handwritten characters than for the machine-printed digits. This was apparently because some of the students in filling out the cards did not write dark enough. In order to verify that the relatively high error rate was not due to a poor signal-to-noise ratio in the scanner, a set of 60 samples of each digit having no zero counts in the central region of the character was selected. (A zero count in the center would indicate a lightly written number.) These 60 samples were divided in half. Thirty were used to generate the statistical matrix and the other 30 were tested against the matrix. The results of this test showed only slight improvement over the previous test, indicating that the scanner was apparently functioning satisfactorily on the penciled characters. These results were as described in the Appendix (Test 3).

Since the handwritten samples tended to be more or less the same height and width, it was decided to test the importance of normalizing the scan lines at each angle for each character. To do this, a set of scan lines at fixed intervals was arrived at, for each angle, in the following way. The computer was programed to search a large number of cards for all the first crossings count other than zero. The next *N* columns were divided into seven equal groups, and the last columns of groups 1, 2, 3, 4, 5, and 6

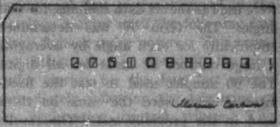


Fig. 6. Sample of handwritten digits used for character recognition test. On the original cards the rectangle surrounding each digit was light blue

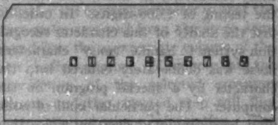


Fig. 7. Sample of machine-printed digits

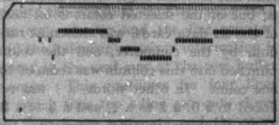


Fig. 8. Sample of data card collected by the experimental scanner for the digit 9 scanned at angle 1

were used as the six data columns, for that angle. The term N was determined empirically for each angle by averaging a large number of cards for all digits. The 60 samples used to test the fixed-interval scan were the same as those used in the previous test. The results were slightly poorer than for the normalized scan method, as can be seen from the Appendix (Test 4).

In the test run on the machine-printed characters, which were essentially non-noisy characters, no errors were made in the testing of 1,000 digits. In order to test the ability of this character recognition system to handle "noisy" characters, noise was deliberately entered into the character by a special program on the computer. The particular kind of noise used here simulated drop-out noise that might occur due to poor printing of the character. This was done as follows: One scan line for each angle of the scan, that is, one of the selected columns on each of the six data cards, was chosen at random by the computer, and the count punched into this column was reduced by one count. In other words, a 1 was reduced to a 0, a 2 to a 1, and a 3 to a 2. These characters in which one-sixth of the data was deliberately entered incorrectly were then tested against the matrix that was generated for use in test 2. The results were as described in the Appendix (Test 5).

Since this system could apparently handle this amount of drop-out noise, the same technique was tried again, but with two scan lines per angle, randomly selected, reduced by one count. These characters in which one-third of the data was deliberately entered incorrectly were then tested against the matrix used in test 2. The results were as described in the Appendix (Test 6).

For the relatively unsuccessful experiment where the statistics used were based on the order of occurrence of the type of crossings, the results obtained are listed in the Appendix (Test 7).

This method, described in Section I, is referred to as the differential crossings count method. For this differential crossings count method, it was noted that, for all characters, the first number of crossings encountered at each angle was almost always a 1. This meant that only two types of crossings were significant at each angle, reducing the useful number of statistics to two-thirds of the total number of statistics processed. To overcome this difficulty, an experiment was run using the first four types of crossings at each of 12 angles, a total of 48 statistics, for a test. The results of this test indicate that this type of crossings information is relatively poor for handwritten character recognition, see Appendix (Test 8).

A Rotating Raster Character Recognizer

To demonstrate the feasibility of instrumenting the rotating raster character recognition technique, a complete transistorized system has been designed. A total of about 250 transistors, 132 cores, and about 1,500 one-per-cent tolerance resistors are required. This section describes the system in detail. It is divided into two parts, one describing the scanning and crossings counting logic and the second describing the analog decision making system.

SCANNING AND CROSSINGS COUNTING LOGIC

The scanning and crossings counting logic operates in the following manner. Before a start pulse is initiated, all triggers so indicated on the block diagram of Fig. 9 are manually reset to the "0," or low-output, condition, and the on-off switch is turned on. The fast-sweep waveform and fast-sweep flyback pulses, items 2 and 3 on the fast-sweep timing chart, Fig. 10, are free-running. This combination would actually consist of a

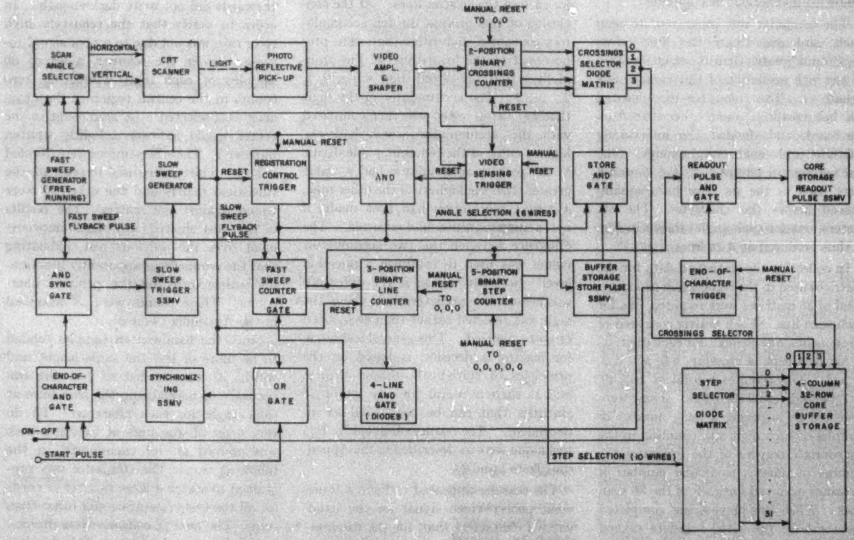

MANUAL RESET TO 0,0
SCAN ANGLE SELECTOR
HORIZONTAL
VERTICAL
CRT SCANNER
LIGHT
PHOTO REFLECTIVE PICK-UP
VIDEO AMPL. & SHAPER
2-POSITION BINARY CROSSINGS COUNTER
CROSSINGS SELECTOR DIODE MATRIX
0
1
2
3
RESET
MANUAL RESET
FAST SWEEP GENERATOR (FREE-RUNNING)
SLOW SWEEP GENERATOR
RESET
REGISTRATION CONTROL TRIGGER
AND
RESET
VIDEO SENSING TRIGGER
MANUAL RESET
STORE AND GATE
READOUT PULSE AND GATE
CORE STORAGE READOUT PULSE SSMV
FAST SWEEP FLYBACK PULSE
SLOW SWEEP FLYBACK PULSE
ANGLE SELECTION (6 WIRES)
AND SYNC GATE
SLOW SWEEP TRIGGER SSMV
FAST SWEEP COUNTER AND GATE
3-POSITION BINARY LINE COUNTER
MANUAL RESET TO 0,0,0
5-POSITION BINARY STEP COUNTER
BUFFER STORAGE STORE PULSE SSMV
END-OF-CHARACTER TRIGGER
MANUAL RESET
RESET
MANUAL RESET TO 0,0,0,0,0
CROSSINGS SELECTOR
0
1
2
3
END-OF-CHARACTER AND GATE
SYNCHRONIZING SSMV
OR GATE
4-LINE AND GATE (DIODES)
STEP SELECTOR DIODE MATRIX
0
1
2
31
4-COLUMN 32-ROW CORE BUFFER STORAGE
ON-OFF
START PULSE
STEP SELECTION (10 WIRES)

Fig. 9. Rotating raster character recognizer scanning and crossings counting logic

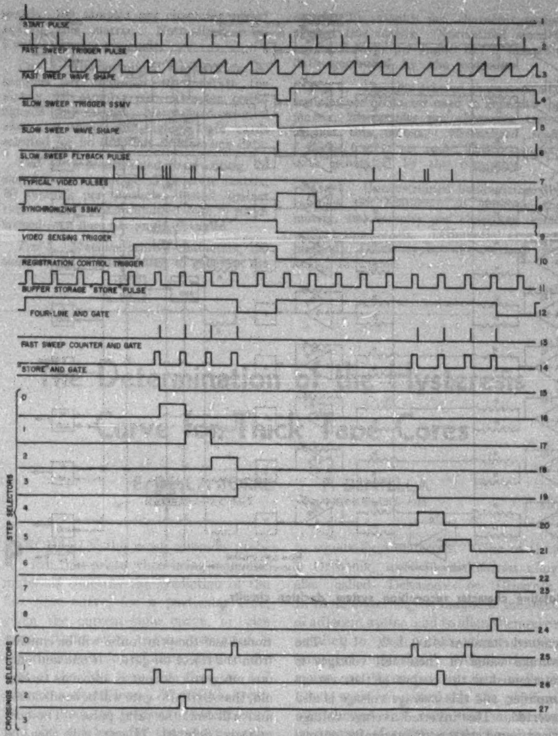


Fig. 10. Fast sweep rotating raster character recognizer logic timing chart

free-running blocking oscillator for the flyback pulses and a linear sweep initiated by these pulses of duration of one half the period of the blocking oscillator. The repetition rate of these pulses is 80 kc. This rate will give a 1-kc character repetition rate. The end-of-character AND-gate is now conditioned to transmit pulses from the synchronizing SSMV (single shot multivibrator) to the AND synchronous gate.

When a start pulse, item 1 of the timing chart, is initiated, the synchronizing SSMV is triggered, emitting a pulse that is slightly longer than the period of the fast sweep period (see item 8). This pulse conditions the AND synchronous gate so that the next occurring fast-sweep flyback pulse can trigger the slow-sweep trigger SSMV. This is shown in item 4 of the timing chart. The slow-sweep trigger SSMV drives the slow-sweep generator which in turn puts out the slow-sweep wave shape, item 5, and the slow-sweep flyback pulse, item 6. The fast- and slow-sweep voltages are mixed in the scan line selector so that the correct faster angle and size are generated on the crt face.

No further changes take place as the raster scans the crt face until the crt spot first crosses the character to be recognized. When the light spot crosses the character, the photoreflective pick-up emits a pulse which is amplified and shaped. This first video pulse triggers the video sensing trigger and is counted by the crossings counter. The video sensing trigger conditions the AND-gate connected to its output so that the next occurring fast-sweep flyback pulse triggers the registration control trigger. This trigger conditions the fast-sweep counter AND-gate and the store AND-gate. This sequence is shown in Fig. 10, items 7, 9, and 10.

The number of crossings encountered during each fast sweep is stored in the crossings counter until the following fast sweep flyback pulse occurs. This crossings information is stored in the core buffer storage for the four fast sweeps immediately following the triggering of the registration control trigger. This sequence is as follows, see Fig. 10, items 10, 14, and 15 through 23. The registration control trigger is triggered when the crossings counter is reset. The crossings encountered on the following fast sweep are stored in the crossings counter and decoded by the crossings selector so that the store pulse received from the store AND-gate drives the appropriate column of the core buffer storage. The core at the intersection of the selected crossings column and the selected step will flip, storing the crossings information. The step row selection is determined from the step counter decoded by the step selector.

The next fast-sweep flyback pulse results the crossings counter, steps both the step counter and the line counter, and the following crossings information is stored in the next row of the core buffer storage. This continues for four lines at which time the 4-line AND-gate closes the fast sweep AND-gate and the store AND-gate. Nothing further occurs until the end of the slow sweep.

At the end of the slow sweep the slow-sweep flyback pulse occurs (see items 5 and 6). The slow-sweep flyback pulse resets the registration control trigger and and the video sensing trigger, and triggers the synchronizing SSMV. On the next fast-sweep flyback pulse the complete sequence as described above is repeated. This complete sequence occurs eight times until step 31 of the step selector is selected. This triggers the end-of-character trigger which closes the end-of-character AND-gate and conditions the readout pulse AND-gate so that the next flyback pulse triggers the core storage readout pulse SSMV, reading the core buffer information out of all rows in parallel. A manual reset pulse followed by a start pulse will cause the complete character recognition cycle to be repeated.

DECISION CIRCUITRY

Assuming a character has been scanned and the crossings information is stored in the core buffer storage, the following is a description of the decision circuitry, Fig. 11, beginning with a core buffer readout pulse.

The readout winding is not shown on the diagram of the cores. This winding threads all cores in series in the same direction. The readout pulse, when it occurs, simultaneously resets all cores that have

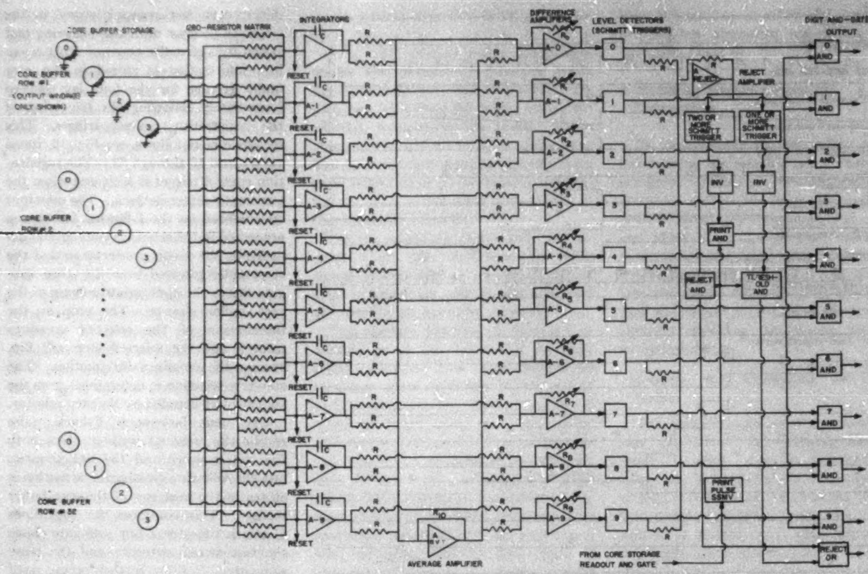

CORE BUFFER STORAGE
1280—RESISTOR MATRIX
INTEGRATORS
CORE BUFFER ROW #1
(OUTPUT WINDINGS ONLY SHOWN)
CORE BUFFER ROW # 2
CORE BUFFER ROW # 32
0
1
2
3
RESET
C
R
A-0
A-1
A-2
A-3
A-4
A-5
A-6
A-7
A-8
A-9
DIFFERENCE AMPLIFIERS
R_0
R_1
R_2
R_3
R_4
R_5
R_6
R_7
R_8
R_9
R_{10}
A AVT
AVERAGE AMPLIFIER
LEVEL DETECTORS (SCHMITT TRIGGERS)
A REJECT
REJECT AMPLIFIER
TWO OR MORE SCHMITT TRIGGER
ONE OR MORE SCHMITT TRIGGER
INV
PRINT AND
REJECT AND
THRESH-OLD AND
PRINT PULSE SSMV
FROM CORE STORAGE READOUT AND GATE
DIGIT AND-GATES OUTPUT
AND
REJECT OR

Fig. 11. Rotating character recognition system, decision circuits

been set during the scanning. There will always be one and only one core set in each row, or 32 in all out of the total of 128. Each core drives ten resistors, one resistor connected to each of ten integrators. The conductance of these resistors is proportional to the logarithm of the matrix probability corresponding to the crossings value, 0, 1, 2, or 3, and the character represented by the integrator, 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. There are 40 resistors for each row giving 40×32, or 1,280, resistors in the resistor matrix. The value of the conductances of these resistors depends on the particular type of font being recognized; typical values would range from about 10^{-4} mhos to 10^{-6} mhos. On the diagram only the resistor connections for the first row of cores is shown; the other rows of cores are connected to the ten amplifiers in the same manner.

When the core buffer readout pulse occurs, the "set" cores flip, emitting a pulse from the output windings. The integrator outputs will be proportional to the sum of the integrals of the core pulse voltage times the conductance of the associated matrix resistors. This means that these integrator output voltages will be proportional to the logarithms of the probabilities that the scanned character is a 0, 1, 2, … 9. The average value of these ten voltages is determined at the output of the average amplifier, and this average voltage is also inverted. This inverted average voltage is compared with each integrator output in the difference amplifiers. Any integrator output that is sufficiently above the average will trigger the associated level-detector Schmitt Trigger. The threshold required for each digit is adjusted by adjustable resistors R_0, R_1, … R_9. If no integrator output is sufficiently high and no level detector is triggered, then none of the digit AND-gates will be conditioned. Then the one-or-more Schmitt Trigger will not be fired and the "threshold AND-gate" will remain conditioned and the print pulse will be emitted from the reject OR-gate. If one and only one integrator output is above its threshold, that digit AND-gate will be conditioned and will emit the print pulse. The one-or-more Schmitt Trigger will then be triggered, closing the threshold AND-gate. If two or more level detectors are triggered, both the one-or-more and the two-or-more Schmitt Triggers will be triggered. This will close the print AND-gate and the threshold AND-gate, and condition the reject AND-gate, causing the print pulse to be emitted from the reject OR-gate. All of the above conditioning takes place during the core storage readout pulse, and the print pulse is initiated by the trailing edge of the readout pulse.

Table I. Computer Decisions in Per Cent

Test No.	Correct	Reject	Error
1	92.0	0	8.0
	88.5	9.5	2.0
2	100.0	0	0
3	93.0	0	7.0
	90.4	8.6	1.0
4	91.0	0	9.0
	86.0	12.0	2.0
5	100.0	0	0
6	94.5	0	5.5
	91.5	8.5	0
7	67.8	0	32.2
8	71.5	0	28.5

Appendix. Results of Tests Performed with the IBM 650 Computer (See Table I)

Test 1. Unconstrained handwriting, 300 samples per character to generate the matrix, 100 samples per character tested against this matrix. Normalized scan,* 6 angles, 6 scans per angle.

Test 2. Machine-printed characters, 100 samples per character to generate the matrix, 100 samples per character tested against this matrix. Normalized scan, 6 angles, 6 scans per angle.

---

* The term normalized scan means that six scan lines per angle were distributed uniformly across the digit as explained previously.

Test 3. Unconstrained handwriting, 30 samples per character to generate the matrix, 30 samples per character tested against this matrix. Normalized scan, 6 angles, 6 scans per angle. Data cards selected for no drop-out in crossings data.

Test 4. Unconstrained handwriting, 30 samples per character to generate matrix, 30 samples per character tested against this matrix. Nonnormalized scan, 6 fixed-spaced scan lines per angle, 6 angles.

Test 5. Machine-printed characters, 100 samples per character used to generate the matrix, 10 samples per character tested against this matrix. Normalized scan, 6 angles, 6 scans per angle, 16.7% incorrect data generated by the computer.

Test 6. Machine-printed characters, 100 samples per character used to generate the matrix, 20 samples per character tested against this matrix. Normalized scan, 6 angles, 6 scans per angle, 33.3% incorrect data generated by the computer.

Test 7. Unconstrained handwriting, 300 samples per character to generate the matrix, 100 samples per character tested against this matrix. Differential crossings method. Three types of crossings per angle, 12 angles.

Test 8. Unconstrained handwriting, 300 samples per character to generate the matrix, 100 samples per character tested against this matrix. Differential crossings method. Four types of crossings per angle, 12 angles.

References

1. SEMINAR ON DECISION THEORY, N. M. Abramson. *Research Memorandum RJ-MR-8*, International Business Machines Corporation, San Jose, Calif., Jan. 1958, 45 pp.

2. ROTATING RASTER CHARACTER RECOGNITION SYSTEM, R. W. Weeks. *Technical Report 16.10.076.005*, Advanced Systems Development Division, International Business Machines Corporation, Dec. 1959, *Appendix* by N. M. Abramson, pp. 27–32.

The Determination of the Hysteresis Curve for Thick Tape Cores

E. DELLA TORRE
NONMEMBER AIEE

P. DENTELLA
NONMEMBER AIEE

PROBABLY the most common problem that arises when using magnetic materials concerns the prediction of the flux–time curve, for a particular core, given the current–time curve, or vice versa. Quite often, the magnetization is a periodic function of time, even in the case of the "d-c loop," so that the ϕ–I curve is a closed loop. Therefore, it can be said that the original problem is to find the ϕ–I loop for the above conditions.

Aside from the type of magnetic material and the peak-to-peak value of the excitation, other, less obvious factors affect the shape of the ϕ–I loop. These include waveform and repetition rate of the excitation, and the geometrical parameters of the core.[1]

Because of the complexity in obtaining a theoretical solution, the problem is customarily solved directly by laboratory experimentation. Still, as an understanding of the physical phenomena involved can help to orient more efficiently the steps in the design of a core, this paper will briefly summarize modern theories on the subject and show their application to a specific case.

The scope of the paper will be restricted to square-loop materials. (One of these is Orthonik, a 50-50 nickel-iron alloy, also called Deltamax or Hipernik.) In these materials the magnetic moments of adjacent atoms tend to align themselves parallel to each other, thus creating saturated "domains." Neighboring domains are separated by thin walls or Bloch walls, so-named for the physicist who showed that each wall must store an amount of energy proportional to its area in order to exist.[2]

Theory of the Magnetization of Square-Loop Materials

For square-loop materials there are only two directions of spontaneous magnetization, which are antiparallel, for all the domains. In general, for a tube of flux within the material, any cross-sectional area is composed of several domains; some are magnetized in one direction, and the remainder in the opposite direction. The net flux in the tube is the difference between the contributions of the oppositely magnetized domains. Any change in this net flux is due to the motion of Bloch walls which changes the area distribution between the two types of domains.

Imperfections in the crystal structure act as obstacles to the motion of the walls; there the walls cannot proceed until the applied field exceeds certain critical values. The average of these values will be called the critical field. For square-loop materials the critical field is equal to the coercive force H_c in the d-c loop and the later terminology shall be used to describe both. For small applied fields the walls stretch elastically without moving past the imperfections, which gives rise to a reversible component of flux (ϕ_R). ϕ_R is a monotonic function of the H field at the wall irrespective of the wall's velocity. Since the flux that would exist in a vacuum of this geometry is a linear function of the applied field, this term shall also be included in ϕ_R. For these materials both components of ϕ_R are of the same order of magnitude.

Another reversible component of flux is due to rotation of the magnetization vector in ferromagnetic domains. This occurs if the crystal axis and the field direction are not parallel. For square-loop materials this effect is very small and shall also be included in ϕ_R. These three components of ϕ_R are difficult to separate experimentally but there is no need to, since they are almost negligible.

The main contribution to the change of flux, however, is due to irreversible motion of the walls past the imperfections. As stated above, this does not occur until the H field at the wall exceeds H_c. The H field at the wall is generally smaller than the externally applied H, due to the

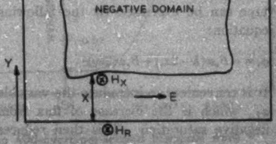


Fig. 1. Cross section of a tape undergoing a change of magnetization

---

Paper 60-200, recommended by the AIEE Basic Sciences Committee for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted January 15, 1958; made available for printing December 23, 1959.

E. DELLA TORRE and P. DENTELLA are with Rutgers University, New Brunswick, N. J.

shielding effect of the eddy currents. This effect is proportional to the velocity of the wall if other conditions are the same.

The motion of the wall is also determined by its inertia. Thus the applied field, less the eddy current field, has to be greater than the coercive force by a term proportional to the wall's acceleration. Since the inertia of the wall is so small, this term shall be neglected.

Analysis of the Magnetization of a Tape-Wound Toroid

To illustrate the concepts cited, a detailed analysis of the magnetization of a tape-wound toroid is now given. In these cores, the thickness h of the ith tape is very small as compared with its width w, and the direction of easy magnetization is made to lie along the circumference.

The H field applied to the surface of the ith tape due to a magnetizing current I is given by:

$$H_r = \frac{NI}{2\pi R_i} \text{ for } R_1 < R < R_2 \tag{1}$$

In this equation N is the number of magnetizing turns, R_i is the radius of the ith tape, R_1 is the inside radius, and R_2 is the outside radius of the core. Fig. 1 shows the cross section of such a tape, whose magnetization is being changed from negative saturation towards positive saturation. The Bloch wall is moving inward from the surface of the tape, thus expanding the volume of the positive domain at the expense of the volume of the negative domain.

The afore-mentioned reversible contribution to the flux can be approximated with sufficient accuracy by

$$\phi_R = kI \tag{2}$$

As shown in Fig. 1, let x be the distance of the Bloch wall from the surface; H_x, the field at the wall; and H_R, the applied field at the surface. If the wall is moving, the flux in the tape is changing, thus inducing an electric field E as shown. If γ is the conductivity of the tape, a current density J equal to $E\gamma$ will result.

Neglecting the effects at the end of the tape, it is seen that the flux ϕ_i of the ith tape can be expressed by the following equation:

$$\phi_i = -B_s w(h - 2x_i) + B_s w(2x_i)$$

It is convenient to introduce the variable ϕ_i', which is the increase of flux from negative saturation. This then reduces to

$$\phi_i' = 4B_s w x_i \tag{3}$$

The rate of change of flux is, therefore,

$$\frac{d\phi_i}{dt} = \frac{d\phi_i'}{dt} = 4B_s w \frac{dx_i}{dt} \tag{4}$$

E can be calculated by taking a line integral in the positive domain, completely enclosing the negative domain.

Following a path parallel to the Block wall, it is seen that

$$\frac{d\phi_i}{dt} = 2wE \text{ for } y < x_i \tag{5}$$

and the corresponding current density is

$$J = E\gamma = \frac{\gamma}{2w}\frac{d\phi_i}{dt} = 2\gamma B_s \frac{dx_i}{dt} \text{ for } y < x_i \tag{6}$$

The dependence of H_x on H_R is given by

$$H_R - H_x = Jx_i \tag{7}$$

If H_R is less than H_c the wall does not propagate, and no currents are induced; then, H_x equals H_R. Conversely, if H_R is greater than H_c, since the field required to accelerate the wall is negligible, the wall must instantaneously move just fast enough so that the eddy currents reduce the value of H_x to H_c. Therefore, for any tape, the following restrictions hold:

$$H_R = 2\gamma B_s x_i \frac{dx_i}{dt} H +_e \text{ if } \frac{dx_i}{dt} \neq 0 \tag{8}$$

Otherwise there is no restriction on H_R.

Rearranging equation 8 and solving for x_i, it is seen that

$$2\gamma B_s \int_0^{x_i} x\,dx = \int_{t_i}^{t} (H_R - H_c)\,dt$$

or, letting $H_R - H_c = H_R'$:

$$x_i = \sqrt{\frac{1}{\gamma B_s}\int_{t_i}^{t} H_R'\,dt} \tag{9}$$

The limits on equation 9 are that x_i must remain less than or equal to $h/2$ and that H_R' either be nonnegative or the integral be set equal to zero during the interval in which it is negative. The limit t_i is the time at which the ring starts to reverse its magnetization. In order to visualize the ϕ–I curve, it must be remembered that the total flux in the core is the sum of all the fluxes in the individual tapes.

Thus, if n is $R_2 - R_1/h$, it follows that

$$\phi' = \sum_{i=1}^{n} 4B_s w x_i \tag{10}$$

Using the value of x_i obtained in equation 9, equation 10 can be further rewritten:

$$\phi' = 4B_s w \sum_{i=1}^{n} \sqrt{\frac{1}{\gamma B_s}\int_{t_i}^{t} H_R'\,dt} \tag{11}$$

In the summation in equation 11 each of the terms is the increase of flux at a different radius tape. If the tape is thin, it can be assumed that the variation is continuous and thus the summation may be reduced to an integral:

$$\phi' = \frac{4B_s w}{h}\int_{R_1}^{R_2} \sqrt{\frac{1}{\gamma B_s}\int_{t_i}^{t} H_R'\,dt}\,dR \tag{12}$$

Expressing this in terms of I, letting

$$f = 4w\sqrt{\frac{B_s}{\gamma}}$$

and

$$b = \frac{N}{2\pi}$$

so that

$$\frac{NI}{2\pi R} = \frac{bI}{R}$$

it is seen that equation 12 reduces to

$$\phi' = \frac{f}{h}\int_{R_1}^{R_2} dR \sqrt{\int_{t_i}^{t}\left(\frac{bI}{R} - H_c\right)dt} \tag{13}$$

Hysteresis Loop for an Applied Triangle Wave

It is now possible to assume an applied current waveform with respect to time and solve for the ϕ–I curve. This will now be done for the simplest case, that is, for a current increasing linearly with time. This is true if a triangle wave of current were applied to the core. Then for some part of the cycle it can be said that

$$I = at \tag{14}$$

The time required to reach a value of H_R equal to H_c depends on the radius; this will be called t_{iR}. This value is

$$t_{iR} = \frac{rH_c}{ba} \tag{15}$$

It is also possible to determine the radius R_i at which H_c is just reached by rewriting equation 15 in this manner:

$$R_i = \frac{abt}{H_c} \tag{16}$$

where R_i will be defined as the radius at which the material is just starting to reverse its magnetization.

In order to find the time t_{fR} at which the ring at radius R is saturated, or to find radius R_f which is saturated at time t, the expression for I of equation 14 will have to be substituted into equations 3 and 9:

$$\phi_R' = f\sqrt{\int_{t_i}^{t}\left(\frac{abt}{R} - H_c\right)dt} = \frac{fH_c}{\sqrt{2ab}}\left[\frac{abt}{H_c\sqrt{R}} - \sqrt{R}\right] \tag{17}$$

If equation 17 is set equal to $2B_s wh$, the saturation flux of one ring, it is then possible to find either t_f or R_f. Then it is seen that

$$(\sqrt{R})^2+p\sqrt{R}-mt=0 \tag{18}$$

where

$$p=\frac{h}{H_c}\sqrt{\frac{B_s ab\gamma}{2}}$$

and

$$m=\frac{ab}{H_c}$$

The solution of this quadratic equation in the square root of R is

$$t_{fR}=\frac{\sqrt{R}}{m}(\sqrt{R}+p) \tag{19}$$

or

$$\sqrt{R_f}=-\frac{p}{2}+\sqrt{p^2/4+mt} \tag{20}$$

In order to perform the integration suggested in equation 13, it is best to break up the integral with respect to radii into three regions, all of which may or may not exist simultaneously. The first is the region in which H_R' is still negative. The contribution to ϕ' of this region is obviously zero, since the flux cannot change until the coercive force is exceeded. This region exists in the interval $R_t<R<R_2$ and no longer exists after R_t has reached the value R_2.

The first region to be discussed which contributes to ϕ' is the region which is completely saturated in the new direction. It exists for $R_1\leq R\leq R_f$ and does not appear until the first ring is saturated. It reaches a maximum value at which it remains when $R_f=R_2$. This contribution to the flux, ϕ_{1f}, is

$$\phi_{1f}=2B_s w(R_f-R_1) \tag{21}$$

If the value of R_f is substituted from equation 20, and the constraints are added, the following is obtained:

$$\phi_{1f}'=2wB_s\left[p^2/2+mt-R_1-p\sqrt{\frac{p^2}{4}+mt}\right] \quad \text{for } t_{1f}<t<t_{f2}$$

$$\phi_{1f}'=0 \text{ for } t<t_{1f} \tag{22}$$

$$\phi_{1f}'=2wB_s(R_2-R_1) \text{ for } t>t_{f2}$$

The remaining region which contributes to ϕ' is the region between R_f and R_t. Here every ring has started reversing its magnetization but no ring has completed it. This contribution will be called ϕ_{ft}'. To find ϕ_{ft}', it is necessary to integrate equation 17 over the region involved; thus

$$\phi_{ft}'=\frac{fH_c}{h\sqrt{2ab}}\int_{R_f}^{R_t}\left[\frac{abt}{H_c\sqrt{R}}-\sqrt{R}\right]dR \tag{23}$$

Upon integration, equation 23 becomes:

$$\phi_{ft}'=2s\sqrt{R}(R/3-mt)\Big|_{R_f}^{R_t} \tag{24}$$

where

$$s=\frac{fH_c}{h\sqrt{2ab}}$$

It must be realized that if $R_f<R_1$ or if $R_t>R_2$, R_1 or R_2, respectively, should be substituted for it. Equation 24 may be rewritten as:

$$\phi_{ft}'=2s[\omega_t-\omega_f] \tag{25}$$

where ω is the value of the right-hand side of equation 24 except for the $2s$, when the corresponding R is introduced.

Using the value of R_t from equation 13, R_f from equation 20, and the values of R_1 and R_2, it is seen that the values of the ω's are

$$\omega_1=\sqrt{R_1}(mt-R_1/3)$$

$$\omega_2=\sqrt{R_2}(mt-R_2/3) \tag{26}$$

$$\omega_t=2/3(mt)^{3/2}$$

$$\omega_f=1/6[p^3+(2mt-p^2)\sqrt{p^2+4mt}]$$

Therefore, the total flux is

$$\phi'=\phi_{1f}'+\phi_t'-\phi_f'+\phi_R \tag{27}$$

where ϕ_{1f}' is given by equation 22, and

$$\phi_t'=\begin{Bmatrix}2s\omega_1 \text{ for } & t\leq t_{t1}\\ 2s\omega_t \text{ for } & t_{t1}\leq t\leq t_{t2}\\ 2s\omega_2 \text{ for } & t_{t2}\leq t\end{Bmatrix}$$

$$\phi_f'=\begin{Bmatrix}2s\omega_1 \text{ for } & t\leq t_{f1}\\ 2s\omega_f \text{ for } & t_{f1}\leq t\leq t_{f2}\\ 2s\omega_2 \text{ for } & t_{f2}\leq t\end{Bmatrix} \tag{28}$$

Experimental Observations

In order to verify the equations a 6-mil tape-wound Deltamax core was tested. It had an inside diameter of 1 inch, an outside diameter of $1^3/_8$ inches, and a tape width of 1/4 inch.

Deltamax has the following physical constants: a resistivity of 51 microhm-centimeters, a saturation flux density of 1.38 webers per square meter, and a coercive force of 16.35 amperes per meter. The last two constants were determined by the method described by Roberts and Van Nice.[1]

With these values, by use of the previous equations, the curve predicted is shown in Fig. 2(A). For comparison,

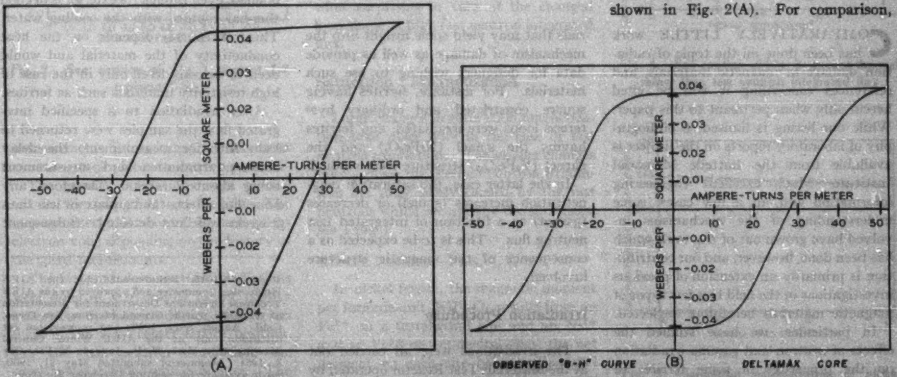


Fig. 2. Comparison of calculated and observed B-H curves for 6-mil Deltamax core due to 400-cycle-per-second triangular wave

A—Calculated curve

B—Observed curve

Fig. 2(B) shows the experimentally observed hysteresis loop for such a core. The similarity between the two shows that the assumptions made were good as a first-order approximation.

Although the results agreed fairly well with the predictions in all cases, it is possible to guess, by using different thickness tapes and different frequency triangle waves, what the probable errors are. The correlation was progressively poorer for ~~thinner tapes~~ (4, 2, and 1 mil). This was probably caused by the degeneration of the simple picture assumed here to a more complex one. Decreasing the frequency probably caused the same result.

The simple assumption that the Bloch wall is a propagating plane is not always valid. It was assumed that its "thickness," or a measure of the roughness of the plane, was negligible compared with the thickness of the tape. This assumption would be less valid for thinner tapes than thicker ones. For thin tapes the forward part of the wall could possibly have met the other wall coming from the other side while the rear part was still relatively close to the surface.

If the applied field is not much larger than H_c, the wall will probably be more ragged than when large fields are applied. The rear part of the wall in the presence of large fields will not be shielded as much by the eddy currents and would move faster than the front and thus try to catch up. For small fields, however, it becomes almost a random process as to which part of the wall moves faster and thus roughness tends to increase.

Nonuniformity of the tapes and a variation of the physical constants of the material as a function of position are other factors which could have accounted for the differences. The first factor does not lend itself to exact mathematical solution. Inclusion of the second factor would greatly complicate the solution. It is believed that both these contributions would be small. In addition, at the frequencies involved, the effect of acceleration is also probably negligible.

When thin tapes are used usually few turns are wound. In this case the inside diameter–outside diameter ratio is approximately 1; the solution has been amply discussed elsewhere.[3] Since in practice cores with large cross-sectional areas are seldom wound with thin tapes, the case where many turns of thin tapes are wound is seldom encountered.

References

1. Influence of ID/OD Ratio on Static and Dynamic Properties of Toroidal Cores, **R. W. Roberts, R. I. Van Nice.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, Nov. 1955, pp. 599–607.

2. Zur Theorie des Austauschproblems und der Remanenzerscheinung der Ferromagnetika, **F. Bloch.** *Zeitschrift für Physik*, Berlin, Germany, vol. 74, 1932, p. 295.

3. Motion of a Ferromagnetic Domain Wall in Fe_3O_4, **J. K. Galt.** *Physical Review*, New York, N. Y., vol. 85, no. 4, Feb. 1952, pp. 664–69.

4. Introduction to Solid State Physics (book), **Charles Kittel.** John Wiley & Sons, Inc., New York, N. Y., 1953.

5. Physical Theory of Ferromagnetic Domains, **Charles Kittel.** *Reviews of Modern Physics*, vol. 21, 1949, p. 541.

6. Magnetic Materials (book), **F. Brailsford.** John Wiley & Sons, Inc., 1951.

7. Ferromagnetic Domains (book), **K. H. Stewart.** Cambridge University Press, Cambridge, England, 1954.

8. Ferromagnetism (book), **R. M. Bozorth.** D. Van Nostrand Company, Inc., Princeton, N. J., 1951.

9. Hysteresis Properties of High Permeability Materials, **T. C. Henneberger.** *M.S. Thesis*, Princeton University, Princeton, N. J., 1953.

10. Analysis of Hysteresis Properties in Ferromagnetic Materials, **E. Della Torre.** *M.S. Thesis*, Princeton University, 1956.

Neutron and Gamma Irradiation of Some Square-Loop and Microwave Ferrites

R. W. MOSS
NONMEMBER AIEE

C. F. KOOI
NONMEMBER AIEE

M. E. BALDWIN
NONMEMBER AIEE

COMPARATIVELY LITTLE work has been done on the topic of radiation damage in magnetic materials, and references concerning it will be cited herein only when pertinent to this paper. While our listing is limited, a bibliography of laboratory reports on the subject is available from the Battelle Memorial Institute. Much excellent engineering information and, in a few cases, some understanding of the mechanisms involved have grown out of the work which has been done, however, and our contribution is primarily an extension of previous investigations in the field to a few types of magnetic materials heretofore neglected.

In particular, we have studied the effects of neutron and gamma irradiation on the properties of some ferrites of special interest to the computer and microwave fields. At the same time, an attempt has been made to choose materials that may yield some insight into the mechanism of damage as well as provide data for designers wishing to use such materials. For instance, ferrites having square, constricted, and ordinary hysteresis loops were irradiated, and ferrites having the spinel ($NiFe_2O_4$) and the garnet ($Y_3Fe_5O_{12}$) structure were studied.

In the latter case, the saturation magnetization increases (spinel) or decreases (garnet) as a function of integrated fast neutron flux. This is to be expected as a consequence of the magnetic structure involved.

Irradiation Procedure

The irradiations were all carried out at the Materials Test Reactor operated by the Philips Petroleum Company at Idaho Falls, Idaho. The samples were encapsulated in sealed aluminum containers, which were cooled by the reactor cooling water. Although the temperature of the samples was not monitored during irradiation, it is estimated that 100 C (degrees centigrade) was never exceeded. In this connection, the gamma heat cannot be conducted out rapidly from the interior of the sample unless the sample is small, even though its exterior is in good thermal contact with the cooling water. This, of course, depends on the heat conductivity of the material and would need to be considered only in the case of high-resistivity materials such as ferrites.

After irradiation to a specified integrated flux, the samples were returned to Palo Alto for measurement, the delay between irradiation and measurement being about 2 weeks. Therefore, any annealing effects taking place in less than 2 weeks were not detected. Subsequent

---

Paper 60-1015, recommended by the AIEE Nucleonics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, San Diego, Calif., August 8–12, 1960, and re-presented for discussion only at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted May 31, 1960; made available for printing December 21, 1960.

R. W. Moss is with the Boeing Scientific Research Laboratories, Seattle, Wash., and C. F. Kooi and M. E. Baldwin are with the Lockheed Aircraft Corporation, Palo Alto, Calif.

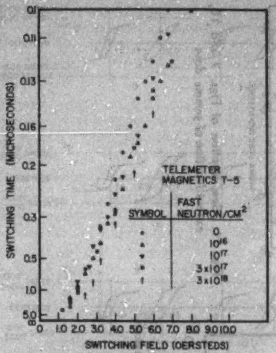


Fig. 1. Switching time as a function of switching field for a manganese-magnesium square-loop ferrite (Telemeter Magnetics T-5) for several integrated fluxes of fast reactor neutrons. The ordinate is in reciprocal scale

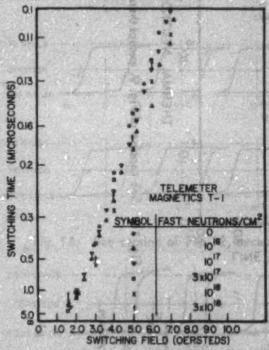


Fig. 2. See caption of Fig. 1, except T-1 replaces T-5

Table I. Neutron-Integrated Flux and Gamma Dose Received by Each Capsule

Capsule Number	Position	Fast Neutron Per Cm²	Slow Neutron Per Cm²	Gamma Radiation, Roentgens
1 (fast neutron irradiation)	Hydraulic Rabbit Facility	10^{16}	0	5×10^{8}
		10^{17}	0	5×10^{9}
		3×10^{17}	0	1.5×10^{10}
		10^{18}	0	5×10^{10}
	L-57 position	3×10^{18}	0	$\sim3\times10^{11}$
2 (slow neutron irradiation)	Hydraulic Rabbit Facility	10^{15}	10^{17}	5×10^{7}
		10^{16}	10^{18}	5×10^{8}
		3×10^{16}	3×10^{18}	1.5×10^{9}
		10^{17}	10^{19}	5×10^{9}
	L-57 position	$\sim10^{18}$	3×10^{19}	$\sim10^{11}$
3 (gamma irradiation)	Spent-Fuel Rod Facility	0	0	10^{7} 3×10^{7} 10^{8} 3×10^{8} 10^{9} 1.1×10^{10}

measurements showed that no long-term annealing took place at room temperature.

The gamma irradiations were carried out in the spent-fuel rod facility. The average energy is about 0.75 mev (meg-electron-volt) depending on the history of the spent-fuel elements.

The fast and slow neutron irradiations were done in a hydraulic rabbit facility except for the last irradiation, which was done in the *L-57* center lattice position. For the fast neutron irradiation, the capsule was shielded with cadmium. The term "fast neutron" as used in this paper refers to neutrons of 1 mev or higher, distributed over a reactor spectrum. The flux of neutrons having energies between 1 mev and thermal neutron energies is not known except by comparison with other reactors.[1] It is probably of the order of 10% of those over 1 mev. Since the neutron dosimetry was good only to ±20%, this error will be neglected. The irradiation received by each capsule is given in Table I.

Switching Speed

The reciprocal of switching time as a function of switching field is plotted in Figs. 1 and 2. The switching time is taken as the time between 10% points on the output voltage pulse. This measurement was made on the same samples of *T-1* and *T-5** as those used in the hysteresis-loop measurement. The data, shown only for the case of fast neutron irradiation, show very little dependence on the integrated neutron flux. This is somewhat surprising in view of the changed loop shape at high fast neutron integrated flux.

Saturation Magnetization

Large changes in saturation magnetization take place only in the case of fast neutron irradiation. The smaller changes taking place for thermal neutron irradiation are most likely caused by the fast neutrons which accompany the slow ones; see Table I.

NICKEL FERRITE—FIGS. 3, 4, 5

In nickel ferrite, the magnetic moment per formula unit, $NiFe_2O_4$, results from an Fe^{3+} on a tetrahedral site and an Ni^{2+} and an Fe^{3+} on octahedral sites, the net moment on each site being antiparallel to the net moment on the other site.[2-4] Since Fe^{3+} has a moment of five Bohr

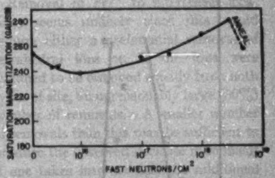


Fig. 3. Saturation magnetization of nickel ferrite as a function of fast neutron integrated flux. Measurements made at field of 3,500 oersteds and temperature of 25 C

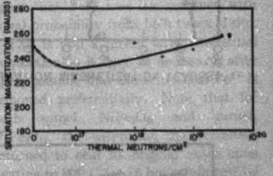


Fig. 4. See caption of Fig. 3, except thermal rather than fast neutron integrated flux

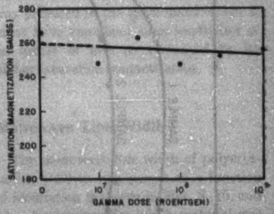


Fig. 5. Saturation magnetization of nickel ferrite as a function of gamma dose. Measurements at same field and temperature as previously

---

* Telemeter Magnetics, Inc., Culver City, Calif.

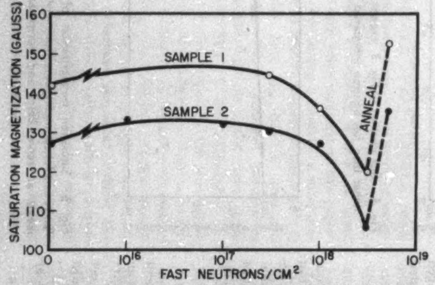


Fig. 6. Saturation magnetization of yttrium iron garnet as function of fast neutron integrated flux. Measurements same as before

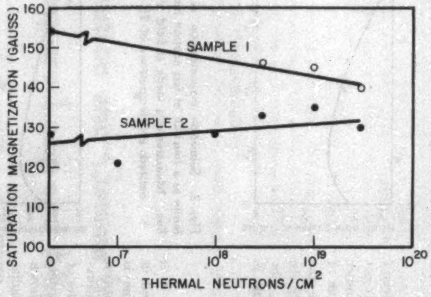


Fig. 7. See caption of Fig. 6, except thermal rather than fast neutron

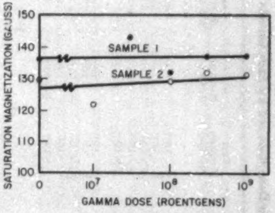


Fig. 8. See caption of Figs. 6 and 7, except function of gamma dose

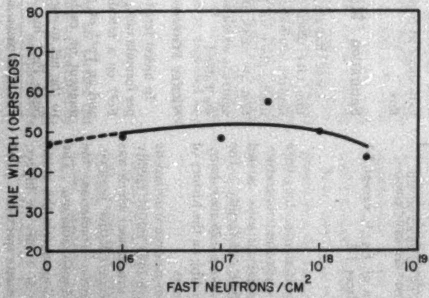


Fig. 9. Microwave line width, at 9,300 mc, of polycrystalline yttrium iron garnet as function of fast neutron integrated flux

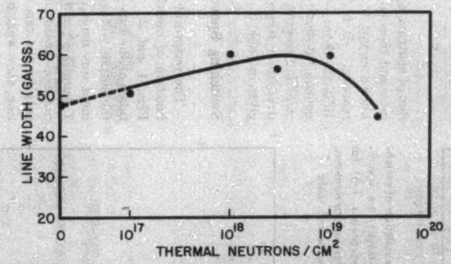


Fig. 10. See caption of Fig. 9, except thermal rather than fast neutron

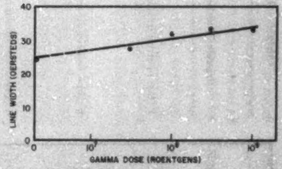


Fig. 11. See caption of Figs. 9 and 10, except function of gamma dose

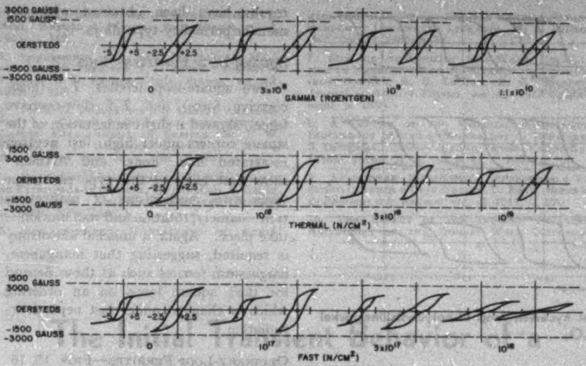


Fig. 12. Effects of gamma, thermal neutron, and fast neutron radiation on the hysteresis loop of permalloy-type ferrite, Ferramic Q

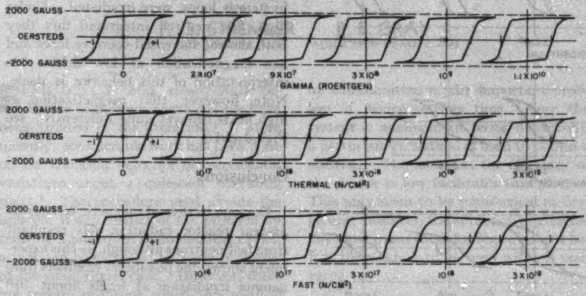


Fig. 13. See caption of Fig. 12, except fast-switching square-loop ferrite (T-1)

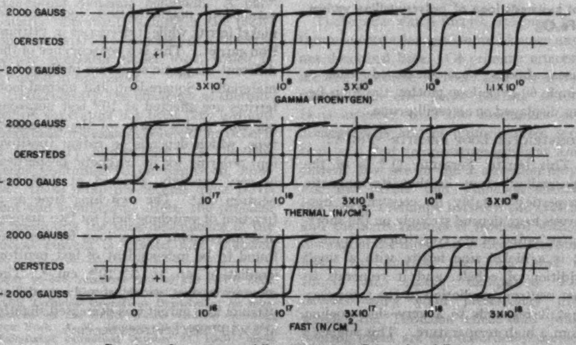


Fig. 14. See caption of Fig. 13, except T-5 replaced T-1

magnetons and Ni^{2+} has a moment of two Bohr magnetons, the transfer of some of the Ni^{2+} to the tetrahedral sites and an equivalent number of Fe^{3+} ions to the octahedral sites will bring about an increase in total magnetic moment. Evidently this has occurred in the case of fast neutron irradiation, but not in the case of gamma and slow neutron irradiation.

This ferrite was prepared by standard ceramic techniques with a final firing at 1,300 C in air.

YTTRIUM IRON GARNET—FIGS. 6, 7, 8

In yttrium iron garnet, the magnetic moment per formula unit, $Y_3Fe_5O_{12}$, results from 3 Fe^{3+} ions on tetrahedral sites and to 2 Fe^{3+} ions on octahedral sites, the net moments of the two types of site being antiparallel.[5] In contrast to nickel ferrite, an exchange of ions between octahedral and tetrahedral sites will not change the magnetic moment. Since a large decrease in magnetic moment occurs in the case of fast neutron irradiation, Fig. 6, a process other than exchange of octahedral and tetrahedral Fe^{3+} ions must occur.

Removal of Fe^{3+} to interstitial locations seems unlikely since this would require either a preferential removal of tetrahedral ions or, if the ions were assumed to be removed equally from both types of site, an unreasonably large (30%) number of removals. A smaller number of removals than this may be sufficient to produce the observed decrease in moment if one takes into account the additional Fe^{3+} ions which will be rendered nonferromagnetic because of incomplete linkages between Fe^{3+} ions, as discussed by Gilleo.[6] With the incomplete linkage mechanism taken into account, only 21% removal of the Fe^{3+} ions would reduce the magnetic moment by 30%, assuming again that Fe^{3+} ions are removed with equal probability from both types of site.

This is still a rather large interstitial population so it may be necessary, after all, to assume that the octahedral Fe^{3+} is removed preferentially. Note that for both spinel $NiFe_2O_4$, and garnet, $Y_3Fe_5O_{12}$, the saturation magnetization returned to near its original value upon heating to 900 C for 16 hours followed by a cool of less than 100 C per hour.

Sample no. 1 was prepared by standard ceramic techniques with a final firing at 1,340 C for 4 hours in air. Sample no. 2 was prepared in the same manner, but it must have contained some impurities or an excess of iron as evidenced by its higher saturation magnetization.

Microwave Line Width

The microwave line width of polycrystalline yttrium iron garnet as a function of irradiation is shown in Figs. 9, 10, and 11. A narrower line width material was used in the gamma irradiation than in the neutron irradiation. The line width at 9,300 mc was measured on spheres of approximately 0.030 inch diameter.

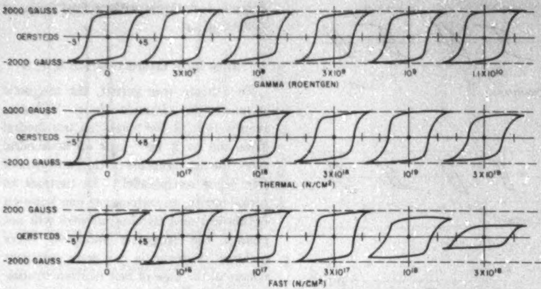


Fig. 15. See caption of Figs. 13 and 14, except hysteresis loop of polycrystalline nickel ferrite $NiFe_2O_4$

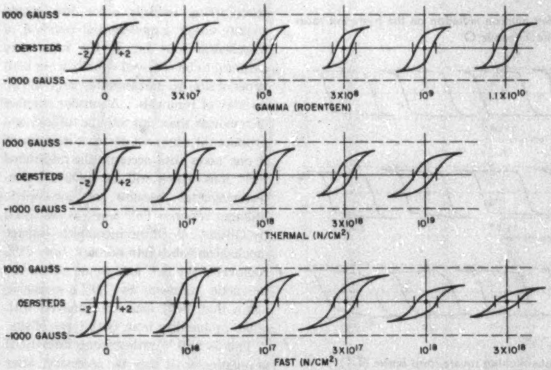


Fig. 16. See caption of Figs. 13, 14, and 15 except hysteresis loop of polycrystalline yttrium iron garnet, $Y_3Fe_5O_{12}$

Although there are no great changes in line width as a function of irradiation, it is surprising that any at all are found, since the line width in polycrystallines such as these is attributable primarily to voids, imperfections, and the granular structure of the material. The line widths of single crystals can be two orders of magnitude smaller.[7] Therefore it may be that drastic broadening of the single-crystal line width occurred. We have no explanation for the decrease of line width at the highest integrated flux for both thermal and fast neutron irradiations.

Hysteresis-Loop Shape

Loop shape was altered appreciably only for the case of fast neutron irradiation. The general tendency was to decrease the magnetization at low fields and to increase the coercive force in some cases. Loops were measured using a simple 60-cycle loop plotter, the loop being displayed on an oscilloscope.

Constricted Loop Ferrite—Fig. 12

This ferrite, Ferramic Q,* is of the "permalloy" type in that its magnetic properties, notably its constricted hysteresis loop, depend strongly on the short-range order of its constituent atoms.[8,9] It is a nickel zinc ferrite with a small addition of cobalt, and it responds to fast neutron irradiation in the same way that it responds to a very slow cooling from a high temperature. This suggests that the same process is operating in both cases. This process is probably a short-range directional ordering along the magnetization in each domain induced by neutron irradiation and resulting in a lower domain wall mobility.[10] Similar results have been obtained with the metallic permalloy-type alloys.[11,12]

Square-Loop Ferrites—Figs. 13, 14

Two square-loop ferrites, *T-1* (high-coercive force) and *T-5* (low-coercive force), showed a slight indentation of the square corners under high, fast neutron integrated flux. Foner and Artman[13] show that loops of this type can occur when two mechanisms of magnetization—namely, rotation and wall motion—take place. Again, a uniaxial anisotropy is required, suggesting that manganese-magnesium ferrites such as these depend for their square loops on an ordering which is changed under fast neutron irradiation.

Ordinary-Loop Ferrites—Figs. 15, 16

Two ferrites, a spinel ($NiFe_2O_4$) and a garnet ($Y_3Fe_5O_{12}$), exhibiting ordinary hysteresis loops, were irradiated. Under high, fast neutron integrated flux they both showed increased coercive force and less magnetization at low fields. No interpretation of this behavior is made. Note, however, that magnetization at high fields behaves quite differently; see Figs. 3 and 6.

Conclusions

Damage to several magnetic ferrites by nuclear reactor radiation was resultant from fast neutrons although, as this experiment shows, the possibility of effects from gamma irradiation at levels about 10^{10} roentgens cannot be ruled out. For fast neutron irradiation of 3×10^{18} cm^{-2}, the saturation magnetization increases in nickel ferrite while it decreases in yttrium iron garnet. This is a consequence of the different magnetic structure of these two materials. Square-loop and normal-loop ferrites are affected at 10^{18} fast neutrons per cm^2 and a ferrite of the permalloy type, which depends on cation (positive ion) ordering for its magnetic properties, is very strongly affected at 10^{18} fast neutron/cm^2. The switching time as a function of switching field for two manganese-magnesium square-loop ferrites was found to be independent of fast neutron irradiation up to 3×10^{18} cm^{-2}. The microwave line width of polycrystalline yttrium iron garnet was decreased slightly at 3×10^{18} fast neutrons per cm^2.

References

1. REDUCTION OF SATURATION MAGNETIZATION OF GAMMA-Fe_2O_3 AND Fe_3O_4 BY PILE IRRADIATION, W. E. Henry, E. I. Salkovitz. *Journal of Applied Physics*, New York, N. Y., vol. 30, 1959 supplement, p. 286S.

2. PROPRIÉTÉS MAGNÉTIQUE DES FERRITES;

---

* General Ceramics Corporation, Keasby, N. J.

Ferrimagnétisme et Antiferromagnétisme, L. Neel. *Annales de Physique*, Paris, France, vol. 3, 1948, p. 137.

3. Conférence sur les Propriétés Magnetique des Ferrites, J. L. Snoek. *Journal de Physique et le Radium*, Paris, France, vol. 12, 1951, p. 228.

4. Physical Properties and Cation Arrangement of Oxides with Spinel Structures, I. Cation Arrangement in Spinels, E. J. W. Verwey, E. L. Heilmann. *Journal of Chemical Physics*, New York, N. Y., vol. 15, 1947, p. 174.

5. The Crystal Structure and Ferrimagnetism of Yttrium-Iron Garnet, $Y_3Fe_2(FeO_4)_3$. S. Geller, M. A. Gilleo. *Journal of Physics and Chemistry of Solids*, New York, N. Y., vol. 3, 1957, p. 30.

6. Superexchange Interaction in Ferrimagnetic Garnets and Spinels Which Contain Randomly Incomplete Linkages, M. A. Gilleo. *Ibid.*, vol. 13, 1960, p. 33.

7. Surface-Independent Spin-Wave Relaxation in Ferromagnetic Resonance of Yttrium Iron Garnet, R. C. LeCraw, E. G. Spencer. *Journal of Applied Physics*, vol. 30, 1959 supplement, p. 185S.

8. A Theory of the Uniaxial Anisotropy Induced by Magnetic Annealing in Ferrites, S. Taniguchi. *Science Reports*, Tohoku University, Sendai, Japan, pt. A, vol. 9, no. 3, 1957, p. 196.

9. Anisotropie Magnétique Superficielle et Substructures d'Orientation, L. Neel. *Journal de Physique et le Radium*, vol. 15, 1954, p. 225.

10. The Effect of the Induced Uniaxial Anisotropy on the Domain Wall Displacements and Magnetic Behavior of Ferromagnetic Cubic Solid Solutions, S. Taniguchi. *Science Reports*, Tohoku University, pt. A, vol. 8, no. 3, 1956, p. 173.

11. Effects of Nuclear Irradiation on Magnetic Materials, R. S. Sery, R. E. Fischell, D. I. Gordon. "Proceedings, Conference on Magnetism and Magnetic Materials," *AIEE Special Publication no. T-91*, Feb. 1957, pp. 453–71.

12. Effect of Neutron Irradiation of the Magnetic Properties and Degree of Order of Magnetic Metal Alloys, A. I. Schindler, E. I. Salkovitz, G. S. Ansell. *Journal of Applied Physics*, vol. 30, 1959 supplement, p. 282S.

13. Magnetization Processes in Heat-Treated Single Crystal Cobalt Ferrite, S. Foner, J. O. Artman. *Ibid.*, vol. 29, 1958, p. 443.

The Initial Transient Behavior of a Thermoelectric Generator

F. G. STREMLER
NONMEMBER AIEE

P. E. GRAY
ASSOCIATE MEMBER AIEE

EFFICIENCY and steady-state behavior of thermoelectric devices have been treated in the literature.[1–3] More recently, considerable work has been done on the transient behavior of incremental variations about a quiescent operating point.[4] The procedure used avoids the inherent nonlinearities of the problem by creating a linearized mathematical model, and it can be extended to many problems of the small-signal case. However, there are certain problems which this method will not handle, the main one of which is large-signal dynamic behavior. A particularly important example is the problem of the start-up transient of a thermoelectric generator under load, which is discussed in this paper.

Until recently, the efficiency of thermoelectric generators has been very low and the details of transient behavior have been of secondary importance. As the efficiency continues to be improved, however, these problems become quite pertinent and significant. Here, the problem of predicting the time-dependence of the output current is treated first, followed by consideration of the important problem of device settling time. Since the system is nonlinear in terms of current, a perturbative method is used to obtain a solution. The fact that the efficiency of the device is low facilitates this process. This may seem to be paradoxical at first glance; however, the efficiency of present-day devices has room for improvement before the method to be used will not converge.

Statement of Problem

The effects which furnish the basis for thermoelectric devices are well-known and not described here. Of primary interest are the Seebeck, Peltier, and Thomson effects, and the Joule heating effect. It is assumed that the conductors are homogeneous and of uniform cross section, and that temperature gradient and current flow are in the x direction only. For the Peltier and Thomson effects, the energy exchange between the system and the surroundings is reversed with a reversal of the direction of electric current flow. In contrast, the Joule heat is simply an energy dissipation. For devices using good present-day materials, the Thomson effect and Joule heating effect are second-order processes for the generator.

A basic device for the thermoelectric generation of power is shown in Fig. 1. The device consists of one p-type semiconductor (labelled as elements a and b for simplicity of notation) to which the electrical and thermal connections are made by metallic contacts. These contacts are assumed, as a first approximation, to furnish ideal conductors to the power sources and sinks. In other words, they are assumed to have negligible specific heat and zero thermal and electrical resistivity. For the practical case, these assumptions can be made quite validly for steady-state operation. However, the first assumption of negligible specific heat cannot be made for the transient case since the thermal mass of the connecting strap is nonzero and influences the transient behavior of the system. The last two assumptions, namely, that the thermal and electrical resistivities are negligible, are retained.

Considerable work can be avoided by choosing convenient notation and normalizing factors at the outset. The parameters of the system are defined as follows:

ρ_a, ρ_b = electrical resistivities of the elements
k_a, k_b = thermal conductivity of the elements
c_a, c_b = specific heats of the elements
A_a, A_b = element cross-sectional areas
τ_a, τ_b = Thomson coefficients of the elements
α_0, α_l = the equivalent differential thermoelectric coefficients at the two junctions
L = specific heat of the hot junction connecting strap

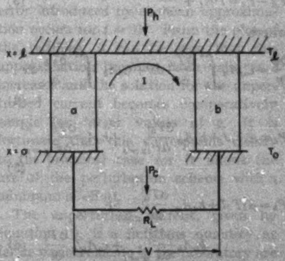


Fig. 1. The thermoelectric generator

Paper 61-225, recommended by the AIEE Solid State Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 27, 1960; made available for printing December 9, 1960.

F. G. Stremler is with the University of Michigan, Ann Arbor, Mich., and P. E. Gray with the Massachusetts Institute of Technology, Cambridge, Mass.

This paper is based on a thesis submitted to the Department of Electrical Engineering of the Massachusetts Institute of Technology in August 1960, in partial fulfillment of the requirements for the degree of Master of Science.

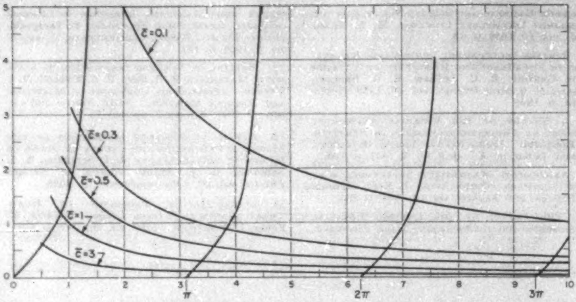


Fig. 2. Graphical solution of the transcendental equation:

$$\tan x = \frac{1}{\bar{c}x}$$

A_o = cross-sectional area of the hot junction connecting strap
l_o = axial length of the connecting strap

It is assumed that the flow of heat is essentially one-dimensional, the elements homogeneous, and the parameters independent of temperature. This last restriction is the most severe and, as shall be shown later, introduces errors into the work. However, as is often the case, the assumptions are required to keep the problem simple enough to yield a tractable solution.

The parameters defined previously can now be used to set up an energy balance for the system. The equations are the conventional ones and need no further comment. For the two elements, a balance of heat flow and stored energy yields:

$$k_aA_a\frac{\partial^2T_a}{\partial x^2}-\tau_aI\frac{\partial T_a}{\partial x}+\frac{\rho_a}{A_a}I^2=c_aA_a\frac{\partial T_a}{\partial t} \tag{1}$$

$$k_bA_b\frac{\partial^2T_b}{\partial x^2}+T_bI\frac{\partial T_b}{\partial x}+\frac{\rho_b}{A_b}I^2=C_bA_b\frac{\partial T_b}{\partial t} \tag{2}$$

For the entire system, the following energy balance applies:

$$P_h-P_c=I^2R_L+\frac{\partial w}{\partial t} \tag{3}$$

where

$$\frac{\partial w}{\partial t}=C_aA_a\int_o^l\frac{\partial T_a}{\partial t}\,dx+C_bA_b\int_o^l\frac{\partial T_b}{\partial t}\,dx \tag{4}$$

$$P_c=\alpha_oT_oI+k_aA_a\frac{\partial T_a}{\partial x}\bigg|_{x=o}+k_bA_b\frac{\partial T_b}{\partial x}\bigg|_{x=o} \tag{5}$$

$$P_h=\alpha_lT_lI+k_aA_a\frac{\partial T_a}{\partial x}\bigg|_{x=l}+k_bA_b\frac{\partial T_b}{\partial x}\bigg|_{x=l}+c_oA_ol_o\frac{\partial T_l}{\partial t} \tag{6}$$

The boundary conditions which are of interest correspond to a device at equilibrium with the surroundings for $t<0$. At a given instant of time ($t=0$, for convenience) a step input of heat is applied to the hot junction ($x=l$) and the cold junction ($x=0$) is kept at the initial equilibrium temperature, T_0. Under these conditions the current output is to be formulated as a function of time under various load conditions. These can therefore be summarized in the following manner:

$$T_a(o,t)=T_b(o,t)=T_o \tag{7}$$

$$T_a(l,t)=T_b(l,t)=T_l(t) \tag{8}$$

$$T_a(x,o)=T_b(x,o)=T_o \tag{9}$$

$$P_h(t)=\delta(t)w \tag{10}$$

where the notation $\delta(t)w$ is used to indicate a step of magnitude w (watts), occurring at the instant $t=0$. These equations form the mathematical description of the thermoelectric generator and hold for any constant parameter system where the system is homogeneous and one-dimensional.

Transients in Unloaded Generator

Before proceeding further, simplification of the preceding equations shall be considered. The simplification which is desirable in equations 1 and 2 is that the terms involving the first derivative of the temperature with respect to the space co-ordinate vanish. This simplifies the differential equations tremendously and, as mentioned previously, corresponds closely to the practical case of the thermoelectric generator for many materials. Therefore, it is assumed that the Thomson heat terms are negligible and consequently $\tau_a=\tau_b=0$ is implied hereafter.

By manipulation of the relations given for the generator, an expression for the current output as a function of the temperature differential across the generator can be found:

$$I(t)=\frac{\alpha T_1(t)}{R_{eq}(m+1)} \tag{11}$$

where

$T_1(t)=T_l(t)-T_o$, R_{eq}

is the effective internal resistance of the generator, and m is the generator load resistance normalized to the effective internal resistance:

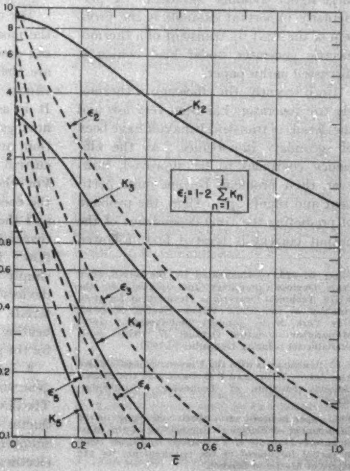


Fig. 3. Residues for the unperturbed solution

$$m=\frac{R_L}{R_{eq}} \tag{12}$$

This result holds for any constant-parameter homogeneous system where the heat flow is assumed to be one-dimensional, the Thomson heat is neglected, and the Seebeck coefficients for the two junctions are equal.

At this point work shall be simplified further by assuming that the parameters of the two arms are very similar so that they can be averaged. In particular, the parameters to be averaged are the Seebeck coefficient, α, the electrical resistivity, ρ, and the thermal conductivity, k. The average is defined as one half of the sum (the arithmetic mean) of the absolute values of each parameter for the two elements. This approximation holds quite well for the case where the two arms are constructed of the same material but with different dopings (for example, p- and n-type bismuth telluride). It simplifies the analysis tremendously since now equations 1 and 2 become identical and only one solution is required. The solution for the case of two dissimilar arms does not require any new theory but it is much more lengthy.

Since the equations considered are nonlinear in terms of the current, a perturbative type of solution is chosen. This method takes advantage of the fact that the efficiency of such a device is rather low so that perturbations to a no-load solution will converge rapidly. As a first approximation then, the temperature distribution for the case of $I=0$ can be obtained, after which equation 11 can be used to find the current that would result from this temperature distribution if there were no other effects. By using this value for the current, a perturbed temperature distribution can be obtained which corrects partially for the initial assumption. With the current set equal to zero, the equation for the temperature distribution becomes:

$$k\frac{\partial^2 T}{\partial x^2}=c\frac{\partial T}{\partial t} \tag{13}$$

which has a solution in the transform domain of the form:

$$T(x,s)=Ce^{\gamma x}+De^{-\gamma x} \tag{14}$$

where

$$\gamma=\left(\frac{c}{k}s\right)^{1/2} \tag{15}$$

Introducing the boundary conditions from the preceding section, and assuming that the thermal conductivity of the connecting strap is high so that the gradient in the connecting strap can be neglected, the matrix equation for the coefficients of equation 14 becomes:

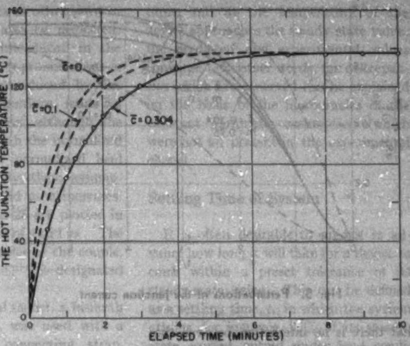


Fig. 4. Unperturbed results for the hot junction temperatures

$$\begin{bmatrix}1 & 1\\ \gamma e^{\gamma l} & -\gamma e^{-\gamma l}\end{bmatrix}\begin{bmatrix}C\\ D\end{bmatrix}=\begin{bmatrix}0\\ \dfrac{w}{2kAs}-\dfrac{t_1}{l}\bar{c}s\,T_1\end{bmatrix} \tag{16}$$

where

$$\bar{c}\triangleq\frac{1}{2}\frac{c_0A_0l_0}{cAl} \tag{17}$$

$$t_1\triangleq\frac{c}{k}l^2 \tag{18}$$

The normalized parameter $\bar{c}$ is the total specific heat of the hot junction connecting strap normalized to the total specific heat of the elements; t_1 (the "characteristic time") is introduced here since it will be apparent later that the time response of the entire system can be normalized on the time scale to the time t_1.

Solving for the coefficients and substituting yields the final result for the temperature variation of the hot junction in terms of the transform variable. The solution in the time domain, which can be determined by the conventional methods of contour integration, is

$$I_1(t/t_1)=I_{1ss}\left[1-2\sum_{k=1}^{\infty}\frac{e^{-\delta_k^2 t/t_1}}{\delta_k^2(\bar{c}\delta_k^2+\bar{c}+1)}\right] \tag{19}$$

where the δ_k's are solutions of the transcendental equation:

$$\tan\delta_k=\frac{1}{\bar{c}\delta_k} \tag{20}$$

and

$$I_{1ss}=\frac{Zw}{\alpha(m+1)} \tag{21}$$

$$Z=\frac{\alpha^2}{\rho k} \tag{22}$$

Note that Z is the conventional figure of merit parameter used in the static analysis of thermoelectric devices.

This is the first of the desired results, namely, the current based upon the unperturbed (no-load) temperature distribution of the device. The form of the solution depends implicitly upon the normalized parameter $\bar{c}$ through the roots of equation 20 and the residues of equation 19. A plot for the roots of equation 20 is given in Fig. 2 for several values of $\bar{c}$, and the values of the residues of equation 19 normalized to a base of 0.500 are given in Fig. 3 as a function of $\bar{c}$. In order to indicate the relative magnitude of the first residue without actually plotting it, the percentage error introduced by breaking off the series of residues at any finite number of poles is shown. The percentage error E is defined at $t=0$, based on the steady-state value of the unperturbed current. With the aid of the graphs of Figs. 2 and 3, the unperturbed current can be written without extensive numerical calculation.

Note that the residues decrease so rapidly that the unperturbed current can be closely approximated by dropping all terms after the first residue, particularly for larger times. The maximum error introduced by such an approximation occurs for $t=0$. From the plot of the residues, it can be seen that this approximation becomes more valid as $\bar{c}$ increases and the solution for the unperturbed current becomes comparatively simple for larger values of $\bar{c}$. It is fortunate that this corresponds closely to the practical case for it allows the use of the perturbation scheme with a minimum of effort.

The unperturbed current, given by equation 19, is a fictitious quantity as far as measurements in the laboratory are concerned, since it describes the current which would flow on the assumption

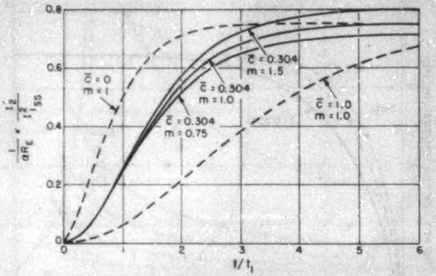


Fig. 5. Perturbations of the junction current

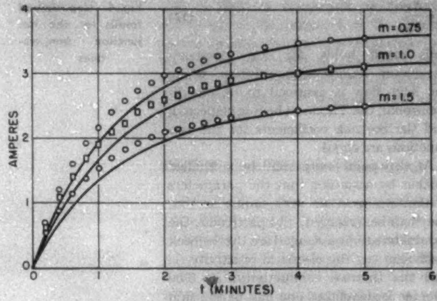


Fig. 6. Comparison of the results with the experiment

that there is no interaction between the temperature distribution and the flow of current. However, an experimental check can be made of the results of this section by monitoring the temperature of the hot junction and checking the experimental results with the theoretical result for the unperturbed temperature.[5] Fig. 4 shows the calculated curve and the measured results for an experimental model which had a measured $\bar{c}$ of 0.304. The predicted temperature dependence for $\bar{c} = 0$ (i.e., neglecting the effect of the structure at the hot junction) is also shown. From these results it is concluded that the value of the normalized parameter $\bar{c}$ has a dominating influence upon the time response of the device for values as small as 0.100.

The Effect of Load on Transient Behavior

The solution of the preceding section neglected the nonlinear effects, and either a perturbation scheme or an iterative process must be used for a final solution. The former method has been chosen because it offers the advantage of calculating the successive corrections to the solution which was obtained and does not require recalculation of the entire solution. Either method must be examined to see if it converges rapidly enough to be used. It is advantageous to remain in the transform domain until the final answer is formulated and then to return to the time domain.

Going back to the equations which first were set up under "Statement of Problem" and then were solved for the special case of $I = 0$ and $\tau = 0$, the equation for the temperature distribution is (for $I \neq 0$, $\tau = 0$):

$$kA\frac{\partial^2 T}{\partial x^2} - cA\frac{\partial T}{\partial t} + \frac{\rho}{A} I^2 = 0 \tag{23}$$

Since an approximate solution for the current has been obtained, it can be substituted into this expression, which can be solved for the perturbed temperature distribution caused by this current. If a matrix equation for the coefficients is set up as in the previous section the result is

$$\begin{bmatrix} 1 & 1 \\ \gamma e^{\gamma_e} & -\gamma e^{-\gamma_e} \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{w}{2kAs} - \frac{l_1}{l}\bar{c}s\, Tl_s \end{bmatrix} - \begin{bmatrix} \frac{\rho}{A^2cs}\beta_1(s) \\ \frac{R_{eq}(m+1)}{2kA}\beta_1(s) \end{bmatrix} \tag{24}$$

Here T_B is the total temperature after the perturbation is taken into account. The perturbation is denoted by primes (so that $T_B = T_B + T_B'$). The term defined as $\beta_1(s)$ is, in the transform domain, the convolution of the unperturbed current with itself or, in terms of the time domain, the transform of the square of the unperturbed current. In this analysis, it is preferable to treat $\beta_1(s)$ in the latter sense since it is easier to see what is happening physically and therefore allows simplifying assumptions to be made in terms of the physical model.

A comparison of these results with equation 15 immediately suggests a perturbation scheme since only the subtractive matrix on the right-hand side of equation 24 has changed with the perturbation. In fact, it is possible to define the coefficients as a set of additive terms, each corresponding to a perturbation matrix on the right. The total solution is then composed of the superposition of these solutions. Such a system might work very well if higher perturbations were desired, although it may become very lengthy if it is necessary to go to the time domain and then back to the transform domain between each successive perturbation. Theoretically, it is possible to appeal to the convolution integral but in practice the equations become too difficult to handle by these methods.

It was noted previously that the residues of the solution for the unperturbed current decrease rapidly for the larger values of $\bar{c}$. It is shown in Fig. 3 that the retention of one residue for $\bar{c} = 0.300$ gives an accuracy loss of 1% at $t = 0$ (which is the point of maximum error) so that, for $\bar{c} = 0.300$ the unperturbed current can be approximated very closely for $t \geq 0$ by the first residue. If this fact is used, the term defined as $\beta_1(s)$ is simplified considerably.

The inversion of the result is lengthy but follows the conventional methods of inversion by contour integration in the transform domain. The perturbation introduces a pole of second order which makes the task of calculating the residues quite tedious. The second-order pole enters since it has been shown in equation 11 that the current and the hot junction temperature are related by constants of the system and therefore have the same time dependence. It follows from this that a second-order pole is introduced when the current–temperature product is taken. In other words, it is the Peltier effect which introduces the second-order pole in the perturbation. The residue of the second-order pole proves quite useful in the calculation of device settling time.

Retaining the method of grouping which follows from setting up the problem in this way, the final result is

$$\frac{I_2'}{I_{1ss}^2} = \frac{\alpha\epsilon(2kA)^{-1}}{m+1}\Bigg\{\Bigg[-(m+1/2) + 2\sum_{k=1}^{\infty}\frac{(m+1+\bar{c})\delta_k \sin\delta_k - 1}{\delta_k^2 g(\delta_k)\delta_k \sin\delta_k} e^{-\delta_k^2 t/t_1}\Bigg] -$$

$$\frac{4}{\delta_1^2 g(\delta_1)}\left[\frac{1}{2}\,\frac{h(\delta_1)+f(\delta_1)\delta_1 \sin \delta_1}{\delta_1^2 g^2(\delta_1)\sin \delta_1}\,e^{-\delta_1^2 t/t_1}-\right.$$
$$2\,\frac{(m+1+\bar{c})\delta_1 \sin \delta_1-1}{g(\delta_1)\delta_1 \sin \delta_1}\,\frac{t}{t_1}\,e^{-\delta_1^2 t/t_1}+$$
$$\left.2\sum_{k=2}^{\infty}\frac{(m+1+\bar{c})\delta_k \sin \delta_k-1}{a(\delta_k)\,\delta_k \sin \delta_k}\,e^{-\delta_k^2 t/t_1}\right]+$$
$$\frac{4}{\delta_1^2 g^2(\delta_1)}\left[-\frac{1}{2}\times\right.$$
$$\frac{\cos\sqrt{2}\delta_1-1+(m+1)\sqrt{2}\delta_1 \sin \sqrt{2}\delta_1}{\delta_1^2[\cos\sqrt{2}\delta_1-\bar{c}\sqrt{2}\delta_1 \sin\sqrt{2}\delta_1]}\times$$
$$e^{-2\delta_1^2 t/t_1}+2\sum_{k=1}^{\infty}\times$$
$$\left.\left.\frac{(m+1+\bar{c})\delta_k \sin \delta_k-1}{b(\delta_k)\delta_k \sin \delta_k}\,e^{-\delta_k^2 t/t_1}\right]\right\} \qquad (25)$$

where the following groups of polynomials have been defined for convenience in numerical computation:

$$a(\delta)=\bar{c}^2\delta^4+(\bar{c}-\bar{c}\,\delta_1^2+1)\delta^2-(\bar{c}+1)\delta_1^2 \qquad (26)$$

$$b(\delta)=\bar{c}\,\delta^4+(1+\bar{c}-2\bar{c}^2\delta_1^2)\delta^2-2\delta_1^2(1+\bar{c}) \qquad (27)$$

$$f(\delta)=2\bar{c}^2\,(m+1)\delta^4-\bar{c}[5\bar{c}^2+(m+3)\bar{c}-2(m+1)]\delta^2-[3\bar{c}^2+(m+6)\bar{c}+(m+3)] \qquad (28)$$

$$g(\delta)=\bar{c}^2\delta^2+\bar{c}+1 \qquad (29)$$

$$h(\delta)=5\,\bar{c}^2\delta^2+3\bar{c}+3 \qquad (30)$$

From this result a test for the convergence of the perturbation process is given by the following:

$$\epsilon(2kA)^{-1}wZ\frac{m+1/2}{m+1}<1 \qquad (31)$$

The degree of this inequality gives some idea of the relative degree of convergence. For the experimental model used, this test of convergence yielded 0.075 for a matched load ($m = 1$). The perturbation process is fully justified for this case. Note that the criteria of convergence are directly related to the efficiency of the device through the figure of merit Z. This result is quite formidable. Fortunately, however, it can be slightly simplified when $\bar{c} \geq 0.300$. This is the condition which justified the approximations leading to this result so that the condition is certainly not restrictive.

In deriving equation 25, the residues at all poles except the first one in the unperturbed solution were neglected. This is also true in the perturbation, and if the residue at a pole is dropped in the unperturbed solution, the residue at the corresponding pole in the perturbation can be neglected and the "correlation" or effect on the residues at other poles in the perturbation can also be neglected. Moreover, the error introduced in the perturbation by these approximations is less than the error introduced by the corresponding approximations made in the unperturbed solution, although the error is a function of both the normalized parameter $\bar{c}$ and the normalized load parameter m. Generally, these assumptions become more valid as $\bar{c}$ increases. The results of equation 25 are plotted in Fig. 5 for several values of $\bar{c}$ and m. The parallel thermal resistance of the couple, which is equal to $\epsilon(skA)^{-1}$, is designated as R_g.

For the experimental model, a bismuth telluride p-n junction was used with a copper heater and connecting strap. The cold junction was kept at ice-water temperature and a power input of 4.15 watts was applied to give a temperature differential of 140 degrees centigrade across the junction under no-load conditions. The load voltage was recorded as a function of time for various loadings of the generator and several runs were made for each loading. The heater was designed to give as small a value of the normalized parameter $\bar{c}$ as was practically possible. The measured value of $\bar{c}$ was 0.304. The parameters of the system used in the calculated results were obtained by using the average values corresponding to the steady-state temperature.

The results of the calculations are compared with the experimental results in Fig. 6 for three different loadings of the generator. The results are fairly consistent; however, they are not in as close agreement as the comparison of the open-circuit hot junction temperature. Note that the discrepancies occur in the short-time region and are consistent from one load condition to another. The results previously obtained indicated that the current output of the device is directly proportional to the effective value of the figure of merit, Z. The experiment indicated that Z varies considerably with temperature, decreasing monotonically as the temperature increases. The over-all effect is that the output current is higher than that which is predicted by the model at lower temperatures, and as the temperature of the device approaches the steady-state value, the predicted and experimental values converge. In other words, the discrepancies which are observed can be explained on the basis of the inaccuracies in the constant parameter assumption (which were not all present in the experimental check).

Settling Time of System

It is often desirable to predict in advance how long it will take for a device to come within a preset tolerance of its steady-state value. This can be defined as a settling time, t_s, in which the system attains, for instance, 95% of its steady-state current output under load conditions. Earlier in this paper, it was shown that both unperturbed solution and perturbation can be closely approximated for times $t > t_1$ for practical values of the parameters. Furthermore, for these longer times, the parameters of the system are not changing very rapidly and can be assumed to be near a "quasi-static" average value.

Making use of the simplifications which were considered in a previous section, all the terms except the constant and the one involving the first residue can be dropped from the unperturbed solution. Referring to this simplified form of equation 19, it is seen (from the convergence criterion) that the perturbation is small compared with the unperturbed current. Also, for long times, the term of the form xe^{-x} dominates the perturbation as long as its residue is sufficiently large in comparison with the residues of the other terms. In accordance with these approximations the approximate long-time function can be written

$$\frac{I}{I_{1ss}}\cong 1-2K_1 e^{-\delta_1^2 t/t_1}-\frac{Z_e(2kA)^{-1}w}{(m+1)^2}\times\left[(m+1/2)-K_{31}t/t_1 e^{-\delta_1^2 t/t_1}\right] \qquad (32)$$

where, from the preceding section,

$$K_1=\delta_1^{-2}g^{-1}(\delta_1) \qquad (33)$$

$$K_{31}=\frac{8[(m+1+\bar{c})\delta_1 \sin \delta_1-1]}{\delta_1^2 g^2(\delta_1)\delta_1 \sin \delta_1} \qquad (34)$$

$$g(\delta)=\bar{c}^2\delta^2+\bar{c}+1 \qquad (29)$$

and δ_1 is the first root of

Table I. Device Settling Times for Constant Power Input

$\bar{c}$	m	Without Perturbation t_s/t_1	With Perturbation t_s/t_1	With Perturbation t_s (Minutes)	Observed Values t_s (Minutes)
0.304	0.75	1.98	1.804	4.23	4.25
0.304	1.00	1.98	1.824	4.14	4.15
0.304	1.50	1.98	1.886	4.28	4.30
1.00	1.00	4.03	3.78		
3.00	1.00	10.0	9.40		

$$\tan\delta_1 = \frac{1}{\bar{c}\delta_1} \qquad (20)$$

A settling time, t_s, can now be calculated on the basis of the time required for the system to come to 95% of the stead-state operation point. Table I shows the results of the calculations for various values of the normalized parameters, and a comparison with the settling time calculated without considering the perturbation terms is indicated.

Device Settling Times for Constant Power Input

It is interesting to see why the settling time of the generator is less with a load than it is for the open-circuit condition. The perturbation results indicate that the interaction of current and temperature distribution does not begin to take place immediately after the initial instant except for very small values of $\bar{c}$. Therefore, the temperature has a chance to build up almost as fast as in the unloaded case. However, the interaction causes the steady-state operating temperature to be less (through the Peltier effect) than that of the open-circuit case so that the transient settles out in a shorter time. Stated differently, the initial rate at which the current transient builds up is approximately equal for the two cases but, under load, the steady-state operating point is lower and the settling time must be less for the loaded generator since there is no overshoot.

It is also interesting to note that the optimum loading ratio on the basis of the settling time is near the point $m = 1$. An exact expression for the optimum value of m is not easy to obtain, however. The solution can be found from a trial-and-error type of solution for each value of $\bar{c}$. It is not clear from physical reasoning why the optimum value should be at a loading ratio of unity.

Conclusions

The results of the solution to the nonlinear problem of predicting the output transients of the thermoelectric generator prove to be quite valid when the solution is obtained in a perturbative manner. It is found that the convergence of such a process is directly related to the device efficiency. Since efficiency is quite low for the materials in present use, one perturbation is sufficient for the practical case. In theory, the methods described here could be applied repeatedly; in practice, however, the results become too unwieldy to handle after the first perturbation. For the materials and operating ranges used at the present time, the methods which have been described offer a convenient and relatively simple approach to the complex problem of predicting transient output and settling time of a thermoelectric generator under load.

References

1. Semiconductor Thermoelements and Thermoelectric Cooling, A. F. Ioffe. Infosearch, Ltd., London, England, 1957.
2. The Efficiency of Thermoelectric Generators, J. M. Borrego, A. H. Lyden, J. Blair. *WADC TN 58-200*, Massachusetts Institute of Technology, Cambridge, Mass., 1958.
3. Calculation of Efficiency of Thermoelectric Devices, B. Sherman, R. R. Heikes, R. W. Ure. *Journal of Applied Physics*, New York, N. Y., vol. 31, Jan. 1960, pp. 1–16.
4. The Dynamic Behavior of Thermoelectric Devices (book), Paul E. Gray. John Wiley & Sons, Inc., New York, N. Y., 1960.
5. The Initial Transient Behavior of a Thermoelectric Generator, F. G. Stremler. *S. M. Thesis*, Massachusetts Institute of Technology, Sept. 1960.

Magnetization and Pull Characteristics of Mating Magnetic Reeds

R. L. PEEK, JR.
MEMBER AIEE

A SEALED REED RELAY embodies one or more sealed contacts, assembled with a coil and with shielding and supporting members. The distinctive feature is the construction of the sealed contact.[1,2] In its simplest form, shown in Fig. 1, a pair of magnetic reeds, the contacting members, are sealed in a glass envelope, and are actuated by the magnetic field induced in them by the coil current.

In another form, there is only one moving reed, the other magnetic member being short and nearly rigid.[3] This construction, using two short members to provide both back and front contacts, has been used with mercury-wetted contact surfaces.[3,4] Both forms have been used in single- and multiple-contact relays, both neutral and with permanent magnet bias used to give polar or locking performance.[5] Recent work has included development of a miniature sealed contact.[6]

To aid in designing sealed contacts and in understanding their characteristics, a study has been made of their performance relations, including dependence of the pull, contact force, speed, and sensitivity on the reed dimensions. While this analysis as presented here concerns the simple case of like-mating reeds, it is, with some modification, applicable to other forms of magnetic-reed contacts.

The major factor controlling performance of sealed contacts is the flux-carrying capacity of the reeds, as limited by the saturation density of the reed material. The treatment given here starts from the relation between the reed flux and the attractive force at the gap, which must deflect the reeds and provide the contact force. This relation is determined essentially by the reed dimensions alone. The sensitivity, as measured by the ampere-turns for operation and release, depends upon the dimensions of the coil and shield-

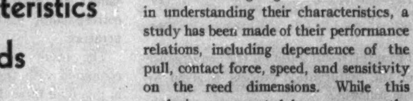


Fig. 1. Dimensions of mating magnetic reeds

Paper 60-1282, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960. Manuscript submitted July 1, 1960; made available for printing November 23, 1960.

R. L. Peek, Jr., is with Bell Telephone Laboratories, Inc., New York, N. Y.

The experimental study described in the section on the pull relation was carried out at Bell Laboratories by R. Calame and Mrs. J. Jones, using apparatus designed and constructed by J. L. Agterberg. The author is also indebted to W. B. Ellwood, O. M. Hovgaard, and other associates for information used in preparing this paper.

ing members, as well as upon those of the reeds. Relations are given for the estimation of reluctance and hence of sensitivity. Finally, expressions are given for estimating the speed of operation, which depends upon the times of field development and of reed motion, with the latter establishing a lower limit to the attainable time of operation.

Pull Relation

Using the notation in Fig. 1 for the reed dimensions, the attractive force F between the reeds is given by Maxwell's law as:

$$F=\frac{\varphi_g^2}{8\pi ab} \tag{1}$$

where φ_g is the flux in the gap where the reeds overlap. The total flux Φ is the sum of φ_g and the fringing flux which passes from one reed to the other by air paths around the gap. As shown subsequently when discussing reluctance, approximate estimates may be made of the fringing flux, but the relations involved are not adapted to a simple direct treatment, relating the pull F to the total flux Φ. A direct experimental study therefore was made to determine the approximate form of this relationship.

Pull and flux measurements were made of four different sets of reeds, having the dimensions listed in Table I. Each set of reeds was assembled with overlap values of 25, 50, and 100 mil-in. (milli-inches), and the pull and flux measured for values of gap x in the range from 1 to 10 mil-in. over the range of coil energization from 0 to 300 ampere-turns.

All measurements were made while the reeds were supported in a brass fixture, which could be adjusted to make the overlap surfaces parallel and to set the overlap and the gap at desired values. The coil had inside and outside diameters of 0.39 in. and 0.86 in., and was 1.64 in. long. It was centered over the gap in making measurements and was provided with a central search coil on its inner diameter. Measurements were also made of the flux in the coil alone; this air-core flux was subtracted from the flux readings to give total reed flux Φ.

Table I. Dimensions of Reeds Tested

Thickness, h (Mil-In.)	Width, b (Mil-In.)	Length (In.)
10	100	1.75
20	100	1.75
30	100	1.75
21	60	1.75

If the reeds are long in comparison with the overlap, as in practice and in these measurements, the pull is independent of reed length. If the permeability of the reed material is high enough for the reeds to be essentially equipotential surfaces near the gap, the pull F is a function of Φ, x, a, b, and h only. For dimensional consistency, therefore, the pull must be given by an equation of the form

$$\frac{8\pi abF}{\Phi^2}=f\left(\frac{x}{a},\frac{a}{b},\frac{h}{b}\right) \tag{2}$$

This is a convenient dimensionless form, as the left-hand term must, from equation 1, approach unity for $x=0$. (In evaluating this term, consistent units must be used: in cgs (centimeter-gram-second) units. Φ is in maxwells, a and b in centimeters, and F in dynes.) From direct plots of the measured values of F and Φ, there were read values of F for $\Phi=10{,}000\ bh$ (i.e., for a flux density of 10,000 gausses). For each set of reeds and each overlap, the corresponding values of $\Phi^2/(8\pi abF)$ were plotted against x/a, as illustrated in Fig. 2, which shows the results for $h=20$ mil-in., $b=100$ mil-in., and three values of a. As the points for all three values of a fall on the same curve, the right-hand side of equation 2 is independent of a/b, and reduces to a function of x/a and h/b only. The experimental relation of Fig. 2 is linear, and similar linear plots, independent of a/b, were obtained for the other three sets of reeds. Thus the functional relation was found to be of the form

$$\frac{\Phi^2}{8\pi abF}=1+k\frac{x}{a} \tag{3}$$

where k is a function of h/b. The observed values of k for the four sets of reeds were plotted against corresponding values of h/b, and a linear plot was obtained, conforming to the equation

$$k=6.66+44.4\frac{h}{b} \tag{4}$$

These results were obtained for values of Φ corresponding to a density of 10,000 gausses. Similar plots were made for other values of density, and substantially the same relation was found to apply for densities ranging from 20% to 80% of the saturation density, which was about 15,000 gausses for the material used in these tests.

Within this range of reed flux density, therefore, all observed values of pull conformed approximately to equation 3, with k given by equation 4. These expressions therefore may be used to estimate the pull of mating magnetic reeds at densities up to 80% of saturaton, when each reed is long as compared with the overlap. The expressions are approximate, and minor deviations, particularly in the value of k, are produced by changes in coil length and in the return path and shielding configuration.

A further and more important deviation from these relations occurs in the closed-gap condition, $x=0$, where the pull is given by equation 3 as $\Phi^2/(8\pi ab)$. For release, this pull equals the retractile force of the reeds, shown in Fig. 3 as sX. Thus, the flux at which release occurs should be given by $\sqrt{8\pi absX}$. Observed values of release flux show much variation, with the upper limit close to the computed value. This corresponds to the fact that surface irregularities or lack of parallelism of mating surfaces concentrate the closed-gap flux over a smaller area than ab, and increase the pull over that given by equation 3 for $x=0$. The actual closed-gap pull therefore is variable, and generally higher than the computed closed-gap pull.

Capability

For operation, the pull curve of the reeds must exceed their stiffness load, shown in Fig. 3 as the force $s(X-x)$. Here, X is the open-gap separation, and s is the combined stiffness—or half the stiffness of one reed if the two are alike. For unequal reeds, $1/s$ is the sum of the compliances of the two reeds. Let F_0' be the closed-gap pull, $\Phi^2/(8\pi ab)$, for the minimum value of Φ required for operation. Then from equation 3, the pull for this value of Φ is

$$F=\frac{F_0'}{1+Cx} \tag{5}$$

where $C=k/a$. At the point of tangency $x=x'$, the pull and load curves must be equal and have equal slopes, so that

$$\frac{F_0'}{1+Cx}=s(X-x')$$

$$\frac{CF_0'}{(1+Cx')^2}=s$$

From these two conditions

$$\frac{x'}{X}=\frac{CX-1}{2CX} \tag{6}$$

and if $CX>1$

$$\frac{F_0'}{sX}=\frac{(1+CX)^2}{4CX} \tag{7}$$

If $CX<1$, the tangency condition no longer applies, the pull and load are equal at $x=0$, and $F_0'=sX$.

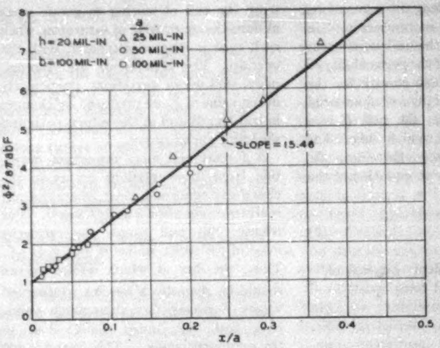


Fig. 2. Results of pull measurements

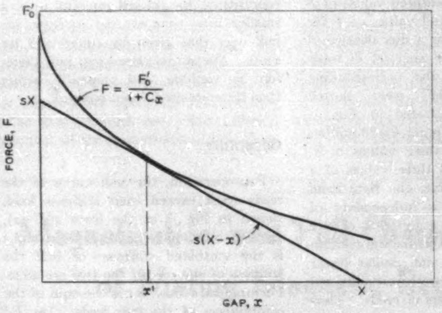


Fig. 3. Pull and retractile force relations

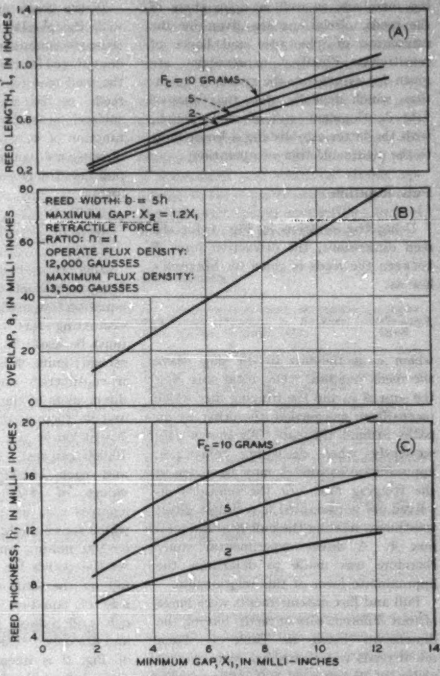


Fig. 4. Illustrative computed reed dimensions

After operation, the flux exceeds the just-operate value. Let F_0 be the value of the closed-gap pull $\Phi^2/(8\pi ab)$ for the final operated flux. Then the contact force is equal to $F_0 - sX$, and the retractile force, tending to open the contact on release, is sX. The contact behavior varies with the contact force, so that one design requirement is that the latter must exceed some specified minimum value F_C.

A second requirement is that the ratio of retractile to contact force must exceed some minimum value n to minimize the danger of sticking. The area of actual intimate contact, and hence the force required to break cold welds over this area, vary directly with the contact force. In formulating this requirement, the operated pull is taken as given by F_0. Since the actual pull is variable and generally exceeds F_0, the actual contact force is usually in excess of its computed value, and the actual value of the retractile force ratio n is less than its computed value.

A satisfactory design must meet the requirements for operation, contact force, and retractile force through the range of dimensional variation occurring in manufacture. The present discussion is confined to the case where only variations in the gap X need be considered. Let the minimum and maximum gap values be X_1 and X_2, and let $c_1 = X_2/X_1$. The final flux must be below saturation, and for estimation purposes may be assumed to correspond to a density B'' of the order of 90% of the saturation density. For operation without excessive reed reluctance, the operating density B' should be of the order of 80% of saturation. Let $c_2 = F_0/F_0' = (B''/B')^2$. Then the operate condition, equation 7, will be satisfied for all values of X if

$$\frac{F_0}{sX_2} = \frac{c_2(1+CX_2)^2}{4CX_2}$$

The contact force requirement is satisfied for all values of X if

$$F_0 = F_C + sX_2$$

The retractile force requirement that $sX > n(F_0 - sX)$ is satisfied for all values of X if

$$\frac{F_0}{sX_1} = \frac{n+1}{n}$$

From the preceding three equations, the design requirements can be written as

$$\frac{4CX_2}{(1+CX_2)^2} = \frac{c_1c_2n}{n+1} \tag{8}$$

$$F_0 = \frac{F_C}{1 - \frac{nc_1}{n+1}} \tag{9}$$

$$s = \frac{F_0 - F_C}{X_2} \tag{10}$$

Reed dimensions must be chosen to satisfy equations 8 through 10 for given values of X_1, c_1, F_C, and n (c_2 being approximately fixed by the reed material).

The dimensions to be selected are a, b, h, and any others which determine s: the free length l, if the reeds are of uniform cross section. If the section ratio b/h is tentatively taken as fixed by manufacturing considerations, then the reed dimensions can be evaluated as follows.

The value of C is computed from equation 8. As $a=k/C$, and k is given by equation 4, the required value of a is thereby determined.

As the value of Φ for F_0 is bhB''

$$F_0=\frac{bh^2B''^2}{8\pi a} \tag{11}$$

With F_0 evaluated from equation 9, and a, B'', and b/h known, the thickness h, and hence the width b, can be determined from equation 11. The stiffness s is given by 10. For a reed of uniform section $b\times h$, the length l can be determined from the required value of s. As equation 11 determines the value of bh^2, and the stiffness varies as bh^3, the shortest reed meeting the requirements will be the one having the largest b/h value consistent with manufacturing considerations.

The outlined relations allow for variations only in the gap X; not in other dimensions. In practice, however, allowance must also be made for variations in the reed thickness h. This can be done by applying a similar treatment to the case where h has upper and lower limits, leading to expressions similar to equations 8, 9, and 10.

Illustrative Computations

Reed dimensions were computed from these relations to give the three values of minimum contact force Fc, shown as parameters for the curves of Fig. 4, over the indicated range of gap dimensions. These illustrative cases were computed for the values of b/h, X_2/X_1, n, and flux density, shown in the legend of this figure. The flux-density values are those applying to iron-nickel alloys of about 50% nickel, which have a Young's modulus value of about 25×10^6 psi (pounds per square inch), which was taken to apply in computing the length required to meet the stiffness requirement.

The computed dimensions in Fig. 4 are reed thickness h, length l', and overlap a. The overlap varies inversely with the gap, as shown by equation 8 when b/h, n, c_1, and c_2 are fixed, as in these computed cases. Both thickness and length increase as the gap and required contact force increase, the gap being the major factor in control of length and, hence, of the sealed contact's over-all size. The choice of gap is partially fixed by voltage breakdown requirements, and partially by manufacturing considerations which determine the variation in gap for which allowance must be made. The comparisons in Fig. 4 are somewhat idealistic in that a constant ratio $c_1(=X_2/X_1)$ is assumed. Actually, ratio c_1 is usually larger for small gaps than for large; in which case reduction in reed thickness and length resulting from reducing the gap is less than that shown in Fig. 4.

In computations for Fig. 4, retractile force ratio n is taken as unity. This is the computed minimum ratio of retractile to contact force. As discussed previously, the actual contact force may be increased by flux concentration in the closed gap, and the actual ratio may be materially below the computed minimum, which serves therefore as a basis for comparing one design with another, rather than as an absolute measure of the liability to sticking. From equation 9, increasing n increases F_0 and, in turn, thickness h and required length l'.

Stiffness Estimation

Reed length denoted as l' in Fig. 4 is that giving the desired stiffness ($2s$ from equation 10) for a uniform reed of length l' and stiffness given by $Ebh^3/(4l'^3)$. As shown in Fig. 1, convenient construction for the seal requires the use of a cylindrical section extending from it for a length cl, with the flat section extending beyond this for a length l. The cylindrical and rectangular sections are of equal area. The stiffness of such two section cantilevers are given in Fig. 6-30 of Peek and Wagar,[7] from which can be derived the expression for ratio l'/l of two cantilevers of equal stiffness, one having the dimensions shown in Fig. 1, and one of length l', and having the same rectangular section as the other.

$$\frac{(l')^3}{(l)^3}=1+\frac{(\pi h)}{(3b)}\left[(1+c)^3-1\right] \tag{12}$$

Thus, if cl and l are known, l' can be evaluated and used to determine the stiffness. Conversely, if l' is determined from the required stiffness, as in Fig. 4, the corresponding flat reed length l can be determined for a given value of cl.

Reluctance for an Air-Return Path

The flux Φ between the reeds required for operation can be estimated by relations given in preceding sections. The corresponding value of coil magnetomotive force $\mathfrak{F}$, or $4\pi NI$, is given by $\mathfrak{R}\Phi$, where $\mathfrak{R}$ is the reluctance of the magnetic circuit. To estimate the sensitivity, or the value of $\mathfrak{F}$ for operation, therefore requires some method of estimating the value of $\mathfrak{R}$.

Most reed relays have essentially an air-return path, whose reluctance is only slightly reduced by the can or shield placed over the core. For an air return, the flux through the reeds is analogous to the current through a leaky transmission line, as discussed in section 9-2 of Peek and Wagar.[7] Referring to Fig. 5, let φ be the flux in the reed at a distance x from the plane of symmetry, and f the potential difference between x and the plane of symmetry. The flux $d\varphi$ leaving the reed over the length dx is given by

$$-pf=\frac{d\varphi}{dx} \tag{13}$$

where p is the permeance, pu (per unit) length of reed of the air path from x to the plane of symmetry. Writing q for the impressed magnetomotive force pu length of coil, $\mathfrak{F}/L_1$, the rate of change in f is given by

$$\frac{df}{dx}=q-r\varphi \quad \text{for } x<\frac{L_1}{2}$$

$$\frac{df}{dx}=-r\varphi \quad \text{for } \frac{L_1}{2}<x<\frac{L_2}{2}$$

where r is the reluctance of the reed pu length. Substituting equation 13 in these expressions gives

$$\frac{d^2\varphi}{dx^2}-pr\varphi+pq=0 \quad \text{for } x<\frac{L_1}{2}$$

$$\frac{d^2\varphi}{dx^2}-pr\varphi=0 \quad \text{for } \frac{L_1}{2}<x<\frac{L_2}{2} \tag{14}$$

Subject to limitations discussed later, p and r will be considered constants, in which case the solution to 14 is given by

$$\varphi=\frac{q}{r}+Ae^{\alpha x}+Be^{-\alpha x} \quad \text{for } x<\frac{L_1}{2}$$

$$\varphi=Ce^{\alpha x}+De^{-\alpha x} \quad \text{for } \frac{L_1}{2}<x<\frac{L_2}{2} \tag{15}$$

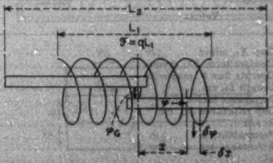
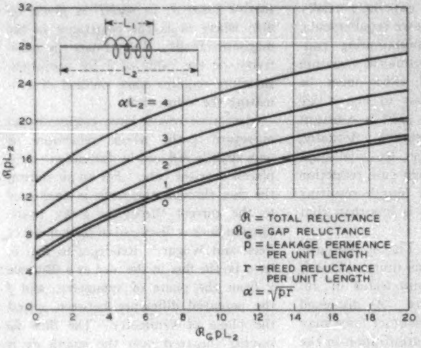


Fig. 5. Notation for analysis of air-return magnetic field



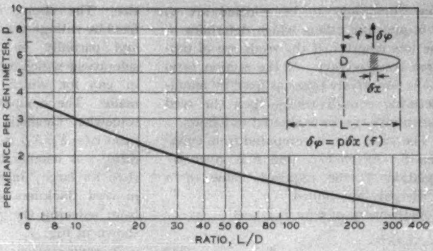


Fig. 6 (left). Reluctance relation for $L_1/L_2 = 0.5$

Fig. 7 (above). Permeance factor for ellipsoidal magnets

where

$$\alpha = \sqrt{pr} \tag{16}$$

For continuity, the expressions for φ and $d\varphi/dx$ at $x = L_1/2$—given by the two forms of 15—must be equal. From these, expressions are obtained for coefficients C and D in terms of A and B. Coefficients A and B then are determined by the boundary conditions, which require that φ must be zero at $x = L_2/2$; and that for a gap flux φ_G at $x = 0$, the potential $f = -\varphi_G \mathcal{R}_G/2$ at $x = 0$, where $\mathcal{R}_G$ is the gap reluctance. Then, from equation 13, the boundary conditions become

$$\varphi = \varphi_G \quad \text{at } x = 0$$

$$\frac{d\varphi}{dx} = \frac{p\mathcal{R}_G\varphi_G}{2} \quad \text{at } x = 0$$

$$\varphi = 0 \quad \text{at } x = \frac{L_2}{2}$$

Substituting these conditions in 15, there are obtained

$$2A = \left(1 + \frac{p\mathcal{R}_G}{2\alpha}\right)\varphi_G - \frac{q}{r} \tag{17}$$

$$2B = \left(1 - \frac{p\mathcal{R}_G}{2\alpha}\right)\varphi_G - \frac{q}{r} \tag{18}$$

$$\varphi_G = \frac{q}{r}\,\frac{\cosh \alpha L_2/2 - \cosh \alpha(L_2 - L_1)/2}{\cosh \alpha L_2/2 + p\mathcal{R}_G/(2\alpha)\sinh \alpha L_2/2} \tag{19}$$

As the boundary condition for $x = 0$ shows, the flux φ in the reed initially increases with x; hence, the point of maximum flux Φ is at some positive value of x where $d\varphi/dx = 0$. Applying this condition to the expression for φ given by equation 15, the maximum flux Φ is given by

$$\Phi = \frac{q}{r} - 2\sqrt{AB}$$

Maximum flux Φ is the total flux between the reeds. Reluctance $\mathcal{R}$, or $\mathfrak{F}/\varphi$, is given by qL_1/Φ. Substituting from the preceding equations, this expression for $\mathcal{R}$ may be written in the form

$$\mathcal{R}pL_2 = \frac{(\alpha L_2)^2}{\frac{L_1}{L_2}1 - \left[\left(1 - \frac{r}{q}\varphi_G\right)^2 - \left(\frac{Q}{2\alpha L_2}\frac{r}{q}\varphi_G\right)^2\right]^{1/2}} \tag{20}$$

where $Q = \mathcal{R}_G pL_2$, and $r\varphi G/q$ is given by equation 19. Thus, ratio $\mathcal{R}pL_2$ is a function of the three ratios L_1/L_2, αL_2, and Q, or $\mathcal{R}_G pL_2$. For the limiting case, where the reed reluctance pu length r is so small that α approaches zero, the hyperbolic functions in 19 may be replaced by their series expansions and the higher powers in α neglected. Then, for $\alpha \to 0$, equation 20 reduces to

$$\mathcal{R}pL_2 = \frac{\dfrac{2(4+Q)L_2}{2L_2 - L_1}}{1 + \dfrac{Q^2(2L_2 - L_1)L_1}{16(4+Q)L_2^2}} \tag{21}$$

Air-Return Reluctance

Fig. 6 shows $\mathcal{R}pL_2$ for the case $L_1/L_2 = 0.5$ plotted against $\mathcal{R}_G pL_2$ for various values of αL_2. The increase of $\mathcal{R}$ with increasing αL_2 measures the effect of the reed reluctance. Aside from the reed reluctance, $\mathcal{R}$ comprises the air-return reluctance in series with the gap reluctance as shunted by the parallel air path across the gap. The curves show how this shunt path limits the increase in $\mathcal{R}$ with increasing $\mathcal{R}_G$.

The effect of increasing L_1/L_2 is to increase $\mathcal{R}$, as can be seen from equation 21, except for large values of $\mathcal{R}_G$, and hence of Q. Thus, reluctance is reduced and sensitivity increased by using a short coil concentrated over the gap.

In deriving the preceding equations, p and r have been taken as constants. The estimation of the air-path permeance of bar magnets is discussed in a classic paper by Evershed,[8] who showed that satisfactory estimates can be obtained by using the constant value of p applying to a uniformly magnetized ellipsoid. This is the function of the ratio L/D shown in Fig. 7. In applying this to mating rectangular reeds, L is taken as the over-all length L_2, and D as $2(b+h)/\pi$.

The assumed constancy of r, the reluctance pu length, is at best an approximation, as r is inversely propor-

Table II. Illustrative Reluctance and Sensitivity Values

Values	Case 1	Case 2
Gap X_1, mil-in	4	10
Contact force F_C, grams	2	10
Operate flux Φ, maxwells	27	140
Length L_2, cm	3.65	8.68
Permeance factor p	1.76	1.73
Gap reluctance $\mathcal{R}_G$, cm^{-1}	1.30	0.38
$\mathcal{R}_G pL_2$	8.34	5.77
Permeability μ	1,000 and 5,000	1,000 and 5,000
Reluctance per cm, r(cm^{-2})	0.450 and 0.090	0.086 and 0.017
αL_2	3.26 and 1.30	3.18 and 1.50
$\mathcal{R}pL_2$	19.0 and 13.8	17.0 and 12.5
$\mathcal{R}$, cm^{-1}	2.96 and 2.15	1.11 and 0.83
Operate ampere-turns	63 and 46	126 and 93

tional to the permeability of the material, which varies with the flux density, so that r varies along the length of the reeds. If, however, r is small, the error introduced by this approximation is minor.

The effect of the permeability on the reluctance and sensitivity, as related to the size and capability of mating reed contacts, is illustrated by the computed results of Table II. Here, dimensions h, b, a, and l were taken as those shown in Fig. 4 for minimum gap values of 4 and 10 mil-in. and contact force values of 2 and 10 grams. The operate flux values, as in Fig. 4, are for a density of 12,000 gausses. The over-all length L_2 was taken as 3.75 times the computed reed length. The permeance p is from Fig. 7. The critical gap reluctance was obtained by adding a 2-mil-in. air-gap allowance for the closed gap to the critical gap x' computed from equation 6.

With $\mathcal{R}_G pL_2$ thus computed for each of these two cases, values of αL_2 were computed for two values of permeability μ, 1,000 and 5,000. With r given by $1/(\mu bh)$, corresponding values of αL_2 were computed, and corresponding values of $\mathcal{R} pL_2$ read from Fig. 6, taking the coil length L_1 as half of L_2. There thus were obtained the values of reluctance $\mathcal{R}$ shown in Table II, and thus finally the operate values, given in abampere-turns by $\mathcal{R}\Phi/(4\pi)$.

In the range of magnetic-reed dimensions and contact and retractile forces illustrated by these computations, αL_2 is small and the reed reluctance minor for μ of the order of 5,000. The materials used for these reeds have permeabilities of 5,000 and higher at densities of the order of 12,000 gausses, or at about 80% of the saturation density. The permeability falls rapidly as the density approaches saturation, and is of the order of 1,000 at about 90% of saturation density. Hence, a design requiring an operating density over 80% that for saturation may have a materially increased reluctance and consequent loss of sensitivity.

Release Reluctance and Sensitivity

As noted in discussing the pull relation, the computed value of flux at which release occurs is given by $\sqrt{8\pi abs X}$. As this is necessarily a lower value of flux than that for operation, the density is lower and the permeability higher than in operation, so that in the release case the reed reluctance is minor, and the preceding equations can be used to estimate the reluctance for the closed-gap condition. As flat metallic pole faces in contact have an effective air gap x_0 of about 2 mil-in., or 0.005 cm (centimeter), $\mathcal{R}_G$ for this case is given by $x_0/(ab)$. Then the abampere-turns for release are given by $\mathcal{R}\Phi/(4\pi)$, where $\mathcal{R}$ and Φ are the values applying in release. As previously stated, the actual release flux is variable, and generally less than its computed value, and there is therefore a corresponding variation in release sensitivity.

Alternative Reluctance Approximations

Analysis of the two preceding sections is useful in providing an understanding of field distribution in the reeds and coil and of the way in which the leakage field shunting the gap varies with the gap, the flux density, and the reed and coil dimensions. It is, however, limited to the case of an air return, and does not apply to a configuration such as that shown in Fig. 8, where sleeve members are used to couple the reeds to the return path provided by the cover. In such cases, the reluctance can be approximately estimated in terms of the lumped constants of the magnetic circuit shown in the figure. A similar treatment can be applied to the air-return case, giving approximate expressions that are simpler than those here developed.

As indicated in Fig. 8, the coil magnetomotive force is taken as developing an air flux φ_A, which affects the inductance only, and a reed flux Φ, the maximum or total flux Φ of the preceding discussion. The path of Φ comprises the reed reluctance $\mathcal{R}_R$ in series with the sleeve gap reluctance $2\mathcal{R}_S$ and the parallel combination of the gap reluctance $\mathcal{R}_G$ and the shunting leakage reluctance $\mathcal{R}_L$. The path, of course, includes the shielding return members, but these are negligible in reluctance compared with the rest of the path.

The reed reluctance can be estimated from the reluctance pu length $r = 1/(\mu bh)$. The sleeve reluctance can be estimated by an approximation obtained from Fig. 9–8 of Peek and Wagar,[7] from which

$$\mathcal{R}_S = \frac{1}{2\pi L_S} \ln \frac{\pi D}{2(b+h)} \tag{22}$$

where L_S is the sleeve length and D its diameter.

As shown in the air-return case leakage reluctance $\mathcal{R}_L$ varies with gap reluctance, and cannot readily be expressed in a simple form. An approximate expression for the parallel combination of $\mathcal{R}_G$ and $\mathcal{R}_L$, however, can be derived from the experimentally observed pull relation given by equation 3. The pull given by this relation must equal that given by the general pull equation for a variable reluctance $\mathcal{R}_X$,

$$F = \frac{\Phi^2}{8\pi}\frac{d\mathcal{R}_X}{dx}$$

If $\mathcal{R}_X$ is the parallel combination of $\mathcal{R}_G$ and $\mathcal{R}_L$, this expression for F/Φ^2 can be equated to that given by 3, and the resulting expression for $d\mathcal{R}_X/dx$ integrated to give the increase in $\mathcal{R}_X$ from its value for $x=0$. This zero-gap value may be taken as $x_0/(ab)$, with $x = 0.005$ cm, as in the Table II estimates of $\mathcal{R}_G$. Thus $\mathcal{R}_X$ is approximately given by

$$\mathcal{R}_X = \frac{x_0}{ab} + \frac{1}{kb} \ln\left(1 + k\frac{x}{a}\right) \tag{23}$$

where, as before, k is given by equation 4.

Then the total reluctance $\mathcal{R}$ for the magnetic circuit of Fig. 8 is given by

$$\mathcal{R} = \mathcal{R}_R + \mathcal{R}_E + \mathcal{R}_X \tag{24}$$

with $\mathcal{R}_X$ given by 23 and $\mathcal{R}_E$ taken as $2\mathcal{R}_S$ from equation 22. This expression can be used to estimate reluctance for other relay configurations, using expressions for $\mathcal{R}_E$ corresponding to the return

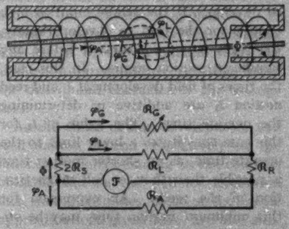


Fig. 8. Magnetic circuit for reeds coupled to a return-path member

Table III. Motion Time Estimates

X_1 Mils	X_1 Cm	F_0, Dynes	s, Dynes per Cm	C, Cm^{-1}	2T, Ergs	m, Grams	t_1, Micro-seconds
4.8	0.0122	5,000	246,000	240	10.2	0.0041	245
12.0	0.0305	25,000	491,000	96	128	0.0549	684

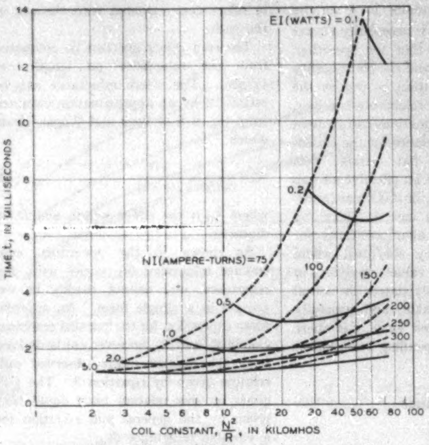


Fig. 9. Observed operate times for 2-switch reed relay

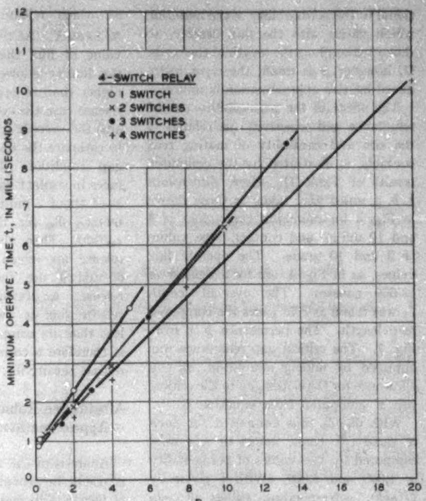


Fig. 10. Observed minimum operate times for reed relay

path coupling and configuration. It may be applied as an approximation to the air-return path case, with $\mathcal{R}_E$ taken as the air-return reluctance. For negligible reed reluctance, this is given by 21 for $\mathcal{R}_G$ and hence Q equal to zero, so that

$$\mathcal{R}_E = \frac{8}{p(2L_2 - L_1)} \qquad (25)$$

While these expressions for the reluctance are approximate, they are simple in form, and are convenient to use, in conjunction with operate and release flux estimates, in estimating operate and release sensitivity.

Motion Time

The operate time of magnetic-reed contacts is that required for field development and reed motion. These two occur together, as reed motion starts as soon as the field and the resultant pull start to develop. If the power input is high enough for the field to develop rapidly, the reeds may be saturated before any significant motion occurs. In this case, the times of field development t_1 and reed motion t_2 are additive in determining the operate time. The value of t_2 for this case constitutes a lower limit to the operate time, as no shorter motion time is possible than that for full magnetization of the reeds. An expression for this minimum motion time may be obtained by the following approximation.

The pull curve is given by equation 3, where Φ has its maximum value, which may be taken as about 90% of saturation, as in the capability estimates. This pull exceeds the just-operate pull shown in Fig. 3, and hence always exceeds the stiffness load $s(X-x)$. The area between the pull curve and the load line represents the kinetic energy supplied to the reeds, which may be denoted $2T$, so that T is the kinetic energy supplied to each of the equal reeds. Writing F_0, as before, for the closed-gap pull for maximum flux, the kinetic energy is given by

$$2T = \int_0^X \left[\frac{F_0}{1+Cx} - s(X-x)\right] dx$$

and hence

$$2T = \frac{F_0}{C}\ln(1-CX) - \frac{sX^2}{2} \qquad (26)$$

While the acceleration is variable, the motion time is nearly the same as that for uniform acceleration, for which the kinetic energy T for each reed is given by

$$T = 2m\left(\frac{X}{2t_2}\right)^2$$

where m is the effective mass of each reed, which moves through half the gap X in time t_2. Thus, the minimum motion time t_2 is given by

$$t_2 = X\frac{m}{2T} \qquad (27)$$

where T is given by equation 26. To illustrate the magnitude of quantities involved for mating magnetic reeds, Table III gives the computed values of t_2 for the two cases of Fig. 4 in which minimum gaps are 4 and 10 mil-in. and contact forces are 2 and 10 grams, respectively. The gap is taken as having its maximum value X_2, and F_0 as corresponding to a density of 13,500 gausses, as before. As the reeds deflect as cantilever beams, the effective mass m of each reed is taken as one quarter of its actual mass, assuming a density of 8.2 grams per cm³. The values of C, or k/a, are those applying for the dimensions given in Fig. 4 for these two cases.

These estimates show that, for the range of reed and gap dimensions covered in Fig. 4, the motion times lie in a relatively narrow band. These times are small compared with those of most other devices for opening and closing contacts in a metallic circuit, but they are large compared with the switching times of many electronic devices.

Operate Time

To estimate the total operate time, an expression is required for the time of flux development. As shown in chapter 4 of Peek and Wagar,[7] an adequate approximation can be obtained from the exponential relation for flux development with the constant reluctance (and hence inductance) for the initial open-gap condition. For magnetic reeds, eddy-current effects are negligible except for flux development so fast that t_1 is negligible compared with t_2. The flux and coil current develop together with a time constant L/R. The time t_1 at which the ampere-turns equal the just-operate value $(NI)_0$ is given by

$$\frac{(NI)_0}{NI}=1-e^{-Rt_1/L}$$

where NI is the steady-state value of ampere-turns. Hence

$$t_1=\frac{L'N^2}{R}\ln\frac{1}{1-v}$$

where $v=(NI)_0/(NI)$ and L' is the single-turn inductance, or $4\pi/\mathcal{R}$, where $\mathcal{R}$ is the open-gap reluctance for the average flux linked by the coil.

If both numerator and denominator in this expression are multiplied by I^2, NI is written as $(NI)_0/v$, and t_2 added to t_1 to give the total operate time t, the resulting expression is

$$t=t_2+\frac{L'(NI)_0^2}{I^2R}\left(\frac{1}{v^2}\ln\frac{1}{1-v}\right) \tag{28}$$

As shown in the text reference cited previously,[7] the function of v appearing in parentheses has a minimum value of 2.5 for $v=0.715$, and departs from this minimum by less than 5% for $0.6<v<0.8$. Thus, for minimum operate time, $NI_0/(NI)$ should lie in this range, and 28 reduces to

$$t=t_2+\frac{2.5nL''(NI)_0^2}{I^2R} \tag{29}$$

where nL'' is written for L'. If there are n sealed contacts (pairs of mating reeds) used in a single relay coil, the single-turn inductance is nL'', where L'' is the single-turn inductance of one sealed contact.

If the total time t is materially larger than t_2, the two terms of equation 29 are not strictly additive, but the approximation has been found to give adequate agreement with experimental observations.

In 28, the term $L'(NI)_0^2$ is proportional to $\Phi(NI)_0$ (where Φ is the operate flux) and, hence, to the field energy required for operation. Therefore, the electric energy input for operation, which is proportional to I^2Rt_1, is proportional to this field energy. As previously noted, equation 28 also shows that t_1 varies with v, or I_0/I, and that the time is a minimum for a given power input if the coil circuit is designed to give I_0/I a value of about 0.7. It also follows from 29 that increasing the steady-state power reduces the operate time until it approaches the lower limit t_2, the minimum time for reed motion.

To use equation 29 for preliminary estimates requires that the operate ampere-turns $(NI)_0$ and the single-turn inductance nL'' be estimated. The former can be estimated by the procedures described in the sections on reluctance estimates. To estimate the single-turn inductance requires estimating the reluctance for the average field linked by the coil. In general, this is a larger reluctance than that for the maximum reed flux, to which the expressions given in preceding sections apply. The maximum flux is limited by reed saturation, and hence controls the capability.

For a tightly coupled magnetic circuit, as in Fig. 8, the maximum and average flux are nearly the same, and the same reluctance expressions apply, except that some allowance may be made for the air field (φ_A in Fig. 8) in estimating the inductance. For the distributed field of an air-path return relay, the average flux linked by the coil is materially less than the maximum value. The analysis of this case given above can be used to give expressions for the average flux, but these are complex in form, and are not included here.

Agreement of experimental observations with equations 28 and 29 is illustrated in Figs. 9 and 10. Fig. 9 shows direct measurements of operate time plotted against the coil circuit constant N^2/R for various values of steady-state power input. The dashed curves are for constant values of ampere-turns NI $[(NI)^2=I^2R(N^2/R)]$. The sealed contacts in this relay had a value of $(NI)_0$ of about 100 ampere-turns, for which the minimum times should occur for a value of NI of about 140, which is near that for the observed minima.

The minimum times for this 2-switch assembly are plotted against $n/(I^2R)$ in Fig. 10, together with those for three other cases with one, three, and four switches. In agreement with 29, the plots are linear and have a common intercept t_2, the minimum motion time. These times were measured to the end of the short chatter interval following initial closure, so that the indicated value of t_2 includes this chatter interval. Slopes of the lines measure the relative value of the single-turn inductance per contact L''; their differences are ascribed to the considerable air-core inductance in this coil, which had space for four switches (sealed contacts). This results in a decrease in the apparent value of L'' as the number of switches increases.

Discussion

The analysis given herein is intended for use in development studies of sealed magnetic-reed contacts. It can be used to estimate the gap and reed dimensions needed to meet design objectives with respect to contact gap, contact and retractile forces, size, speed, and sensitivity. In development studies, models based on the initial estimates can be made and measured, and the measurement results used in conjunction with the analysis as a guide in further work.

In the performance of mating magnetic reeds, the major controlling quantity is the flux between the reeds. This flux, which determines the pull, is limited by saturation and hence by the reed cross section. This total reed flux comprises the gap flux proper and the leakage flux shunting the gap. As shown in the general analysis of the air-return reluctance, this leakage field varies with the gap, reed, and coil dimensions, and the relation between the pull and the total reed flux has a corresponding variation. Hence, the simple pull equation derived from experiment applies rigorously only for the experimental coil length and return-path conditions used in deriving it, and minor deviations occur in applying it to other cases.

Corresponding deviations occur for the reluctance given by the expression derived from the pull equation, and for the other approximate reluctance equations. All these approximations, however, are adequate for preliminary estimates, and may be applied with greater accuracy in detailed studies by using values of the constants that are experimentally determined for the case in question. While the general analysis can be used to estimate the air-return reluctance, and the treatment can be extended to give expressions for the pull and the inductance, it is limited to the air-return case, and is too complex for convenient application. It serves, however, to show the character and extent of the deviations from the simpler approximations resulting from the distributed character of the leakage field.

As the times of flux development and reed motion depend respectively on the inductance corresponding to the coil flux linkages and on the pull for maximum flux, timing estimates, based on the simple approximations given here, are also subject to some minor variation. This can be reduced in development studies by using experimentally determined values of the constants.

The expressions given here apply primarily to mating reeds of equal size, but are approximately applicable to unequal reeds, provided both are long compared with the gap overlap. In such cases, the effective stiffness of the reed combination must be taken as the reciprocal of the sum of the compliances of the two reeds.

References

1. Circuit Closing Devices, W. B. Ellwood. *U. S. Patent no. 2,289,830*, July 14, 1942.

2. Glass Enclosed Reed Relay, W. B. Ellwood. *Electrical Engineering*, vol. 66, Nov. 1947, p. 1104.

3. Mercury Contact Relays, J. T. L. Brown, C. E. Pollard. *Ibid.*, p. 1106.

4. Capability of Sealed Relay Contacts. O. M. Hovgaard. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, Sept. 1956, p. 466–68.

5. Development of Reed Switches and Relays, O. M. Hovgaard, G. E. Perreault. *Bell System Technical Journal*, New York, N. Y., vol. 34, Mar. 1955, p. 309.

6. A Versatile Miniature Switching Capsule, O. M. Hovgaard, W. J. Fontana. *Monograph 3306*, Bell Telephone Laboratories, Inc., New York, N. Y.

7. Switching Relay Design (book), R. L. Peek, Jr., H. N. Wagar. D. Van Nostrand Company, Inc., Princeton, N. J., 1955.

8. Permanent Magnets in Theory and Practice, S. Evershed. *Journal*, Institution of Electrical Engineers, London, England, vol. 58, 1919, p. 780.

Logical Design by Regression

Matrix Logic VII

E. J. SCHUBERT
ASSOCIATE MEMBER AIEE

Synopsis: Catalogs of minimum networks for certain preferred types of building blocks greatly facilitate the logical design. Computational effort in producing such a catalog is modest and depends on the logical functions mechanized in the block. A simple octal coding scheme is introduced for identifying and indexing these functions. Included is a complete catalog of minimum networks for three inputs and one output, assembled from NOR circuits.

ECONOMIC FACTORS favor digital devices designed from a single type of building block. This block must be universal to insure the assembly of any logical network. The criteria for these universal moduli have been previously established. A few symmetric functions related to equivalence have been shown to be the only exceptions to the many universal functions.

Representation of a logical function by its truth table, or equivalent function matrix, lends itself to convenient indexing. This matrix may be regarded as a binary number, which may be replaced by an equivalent one in octal notation. The coding scheme designates logical functions in a catalog, which specifies the inputs required in a single type of modulus in order to obtain a specified function as the output.

To arrive at a logical diagram necessitates table look-up of the output function as designated by its octal index and to note the required input functions for the modulus. These input functions are listed as entries in the same catalog.

Contrary to conventional logical design, this method proceeds from the specified output function toward the initial inputs. This procedure of stepping back from the outputs—one level after the other to the initial inputs—is called logical design by regression.

Permutation Routines

Assuming a switching circuit has p input terminals designated $A_1, A_2 \ldots A_p$, and a single output B, the independent logical variables are given as $X_1, X_2 \ldots X_p$ in addition to 0 and 1. If this circuit implements a universal function, then all possible logical functions of p variables must be obtained by permutation of functions generated previously by the same circuit. Starting with the $p+2$ given initial variables $(0, 1, X_1, X_2 \ldots X_p)$, some new output functions Y_z will be obtained. The routine is exhausted when all possible logical functions of p variables have been generated.

In utilizing the routine for logical design, only the input combination for the FIRST OCCURRENCE of a logical function is recorded; any later occurrence of the same function must be disregarded. Consistent application of this rule yields the MINIMUM LOGIC of single-output nets, comprising one type of universal modulus.

This principle is demonstrated by a convenient example, a 2-input NOR circuit, which performs the operation:

$$B=(A_1+A_2)'=h_1(A_1, A_2) \qquad (1)$$

Operator h_1 is symmetric because transposition of the independent variables at A_1, A_2 does not affect the output B.

Initially available are four of the possible 16 logical functions:

1. Binary signal 0 (0000) = Y_0
2. Binary signal 1 (1111) = Y_{17}
3. Independent variable X_1 (1010) = Y_{12}
4. Independent variable X_2 (1100) = Y_{14}

The matrix notation is already known to be equivalent to a truth table. To denote logical functions Y_z by subscript z, the latter index z is viewed as the octal equivalent of the binary number expressed by the matrix coefficients. To maintain a consistent order of z, the least significant bit in this binary word should correspond to the minterm $X_1'X_2'$.

In Table I, the logical functions are denoted by the index z. Since the initial functions $z = 0, 12, 14, 17$ passed through

Table I. Two-Input Combinations for $h_1(A_1, A_2)=(A_1+A_2)'$

Levels, L =	0	0	0	0	1	1	1	2	2	2	2	3	3	3	3	3
Operands A_1, z =	0	12	14	17	5	3	1	2	10	16	4	15	7	6	13	11
z = 0		5	3	*	*	*	16	15	7	*	13					
12			1	*	*	*	4	*	*	*	*					
14				*	*	*	2	*	*	*	*					
17					*	*	*	*	*	*	*					
Operands A_2 z = 5						10	*	*	6	*	*					
3							*	*	*	*	*					
1								*	*	*	*					
z = 2									*	*	11					
10										*						
16																

* These entries are logical functions which were either initially available or were generated at an earlier step.

Paper 60-1229, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960. Manuscript submitted July 27, 1960; made available for printing November 9, 1960.

E. J. Schubert is a Consultant in Pasadena, Calif.

The author is indebted to the Burroughs Corporation, Research Center, Paoli, Pa., for support of this study during his previous affiliation and, in particular, to Miss R. Baldwin for the assistance in programming the computer Udec III for generation of the catalog presented as part of the paper.

Table II. Catalog of Functions of Two Variables for 2-Input NOR Circuits

Logical Functions z_n:L	Operands A_1	A_2
0.0...Input ZERO		
1.1	12.0	14.0
2.2	1.1	14.0
3.1	0.0	14.0
4.2	1.1	12.0
5.1	0.0	12.0
6.3	1.1	10.2
7.3	0.0	10.2
10.2	3.1	5.1
11.3	2.2	4.2
12.0...Input		X_1
13.3	0.0	4.2
14.0...Input		X_2
15.3	0.0	2.2
16.2	0.0	1.1
17.0...Input.		ONE

no logical element, they are designated by the level $L = 0$. Taken two at a time, they yield six combinations, three of which generate new logical functions denoted by $z = 5, 3, 1$ as the appropriate entries in the table. Because the new functions were obtained after passing through the logical element only once, they are said to be of level $L = 1$.

The process continues until the 16th logical function has been obtained for $z = 11$, at which point all 16 possible functions of the two variables are available.

Computation of Design Catalogs

A suitable computer routine requires a memory location z for each logical function. Inserted in location z is value $L + 1$, the level at which this function became available or first occurred. The so-called initial $p + 2$ logical functions, corresponding to 0, 1, and X_p, are designated by entering in their corresponding memory location z the value for $L + 1 = 1$, because these functions are available at level 0.

Table IV. Logical Design Problem of Defined Function

Variable	Binary Function Table	Octal Index z	Level L
X_1	10 101 010	252	0
X_2	11 001 100	314	0
X_3	11 110 000	377	0
Y	110	6	

Table IV-A. Results of Look-up Catalog

z:L	A	B	C
6.3	0.0	1.1	370.2
370.2	0.0	3.1	5.1
1.1	360.0*	314.0*	252.0*
3.1	360.0*	314.0*	0.0
5.1	360.0*	252.0*	0.0

* These functions at level 0 are initial inputs X_1, X_2, and X_3.

Operator h is represented by a microprogram in the case of h_1 corresponding to the logical sum of A_1, A_2 followed by complementing. Every logical function obtained by this microprogram represents the address z into the memory. If the content of memory location z is $(z) = 0$, then the logical output—representing the result of microprogramming—has not occurred previously. Therefore, the logical functions applied at the inputs for the operands A_1, A_2 have to be recorded together with value L.

As soon as the last reserved memory location has been filled, this subroutine terminates. Sorting the recorded results in the order of index z follows then. Print-out similar to Table II eventually yields the cataloged functions of two variables for a 2-input NOR circuit.

A total of 44 iterations or table entries were required to compose Table I for the symmetric universal NOR function of two variables. The number of possible iterations P are much greater than 44, and might discourage use of the permutation routines for a larger number of input variables. If F denotes the number of all logical functions, the relation is

$$\log_2 F = 2^p \tag{2A}$$

$$\log_2 P \leqq p.2^p \tag{2B}$$

and for $p = 3$

$$\log_2 P \leqq 24; \quad P \leqq 16 \times 10^6 \tag{2C}$$

In an actual computer run for a NOR circuit with three inputs, $p = 3$, approximately 18,500 iterations were sufficient to obtain Table III, the listing of all 256 functions. The ratio of these actual iterations to the number P in equation 2C is less than 0.13%. For $p \geq 3$, a further drastic reduction of this ratio might be expected.

Catalog of Logical Functions

Table III (p. 382), which catalogs the functions of three variables, mechanized by three input NOR circuits, contains four entries in each row. The first column designates the logical function by the octal index z at the left of the period symbol; at the right is the logical level at which this function is obtained by the 3-input NOR circuits.

Using z as reference, logical functions required at input terminals A, B, and C correspond to the operands or logical inputs into the modulus. Designating with Y_L an intermediate logical function at level L, equation 3 describes the recursive routine where, at each function level $L + 1$, a new function has been obtained from at least one earlier logical function at level L. The initial logical functions available at level 0 are listed in equation 4, A through E.

Table V. Logical Design Problem with DON'T CARE Conditions

Variable	Function Matrix	Octal Index z	Level L
X_1	10 101 010	252	0
X_2	11 001 100	314	0
X_3	11 110 000	377	0
Y	−0 −10 011		
	00 010 011	23	3
$Y_{63.1}$	00 110 011	63	1
			MINIMUM
	10 010 011	123	3
	10 110 011	163	3

$$Y_{L+1} = h_1(Y_L, Y_{L-n}, \ldots)\ n = 0, 1\ldots \tag{3}$$
$$Y_{L+1} \neq Y_L$$

$$0 = (00000000) = Y_{0.0} \tag{4A}$$

$$X_1 = (10101010) = Y_{252.0}\ldots 10\ 101\ 010 = 252 \tag{4B}$$

$$X_2 = (11001100) = Y_{314.0}\ldots 11\ 001\ 100 = 314 \tag{4C}$$

$$X_3 = (11110000) = Y_{360.0}\ldots 11\ 110\ 000 = 360 \tag{4D}$$

$$1 = (11111111) = Y_{377.0}\ldots 11\ 111\ 111 = 377 \tag{4E}$$

The corresponding index z is obtained as the octal number equivalent to the binary function matrix. Note that each octal digit corresponds to a group of three bits in the binary word, designating the function matrix. Five initial functions ($p + 2 = 5$ for $p = 3$) are entered under ZERO-level in Table III. The highest octal entry for $z = 377$ designates the logical signal ONE.

Logical design involves the statement

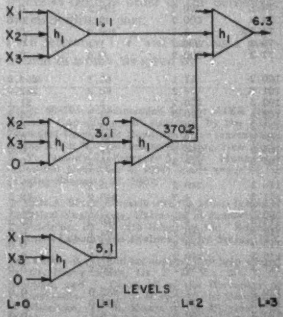


Fig. 1. Logical diagram of NOR circuits designed by regression method

Table III. Catalog of Functions of Three Variables for 3-Input NOR Circuits

z:L	A	B	C
0.0	Input ZERO		
1.1	360.0	314.0	252.0
2.2	125.1	360.0	314.0
3.1	360.0	314.0	0.0
4.2	63.1	360.0	252.0
5.1	360.0	252.0	0.0
6.3	370.2	21.1	0.0
7.3	370.2	0.0	0.0
10.2	63.1	125.1	360.0
11.3	42.2	104.2	360.0
12.2	125.1	360.0	0.0
13.3	104.2	360.0	0.0
14.2	63.1	360.0	0.0
15.3	42.2	360.0	0.0
16.2	21.1	360.0	0.0
17.1	360.0	0.0	0.0
20.2	17.1	314.0	252.0
21.1	314.0	252.0	0.0
22.3	354.2	5.1	0.0
23.3	354.2	0.0	0.0
24.3	352.2	3.1	0.0
25.3	352.2	0.0	0.0
26.3	350.2	1.1	0.0
27.3	350.2	0.0	0.0
30.3	340.2	3.1	5.1
31.3	240.2	42.2	104.2
32.3	340.2	5.1	0.0
33.3	240.2	104.2	0.0
34.3	340.2	3.1	0.0
35.3	300.2	42.2	0.0
36.3	340.2	1.1	0.0
37.3	340.2	0.0	0.0
40.2	17.1	125.1	314.0
41.3	12.2	120.2	314.0
42.2	125.1	314.0	0.0
43.3	120.2	314.0	0.0
44.3	310.2	3.1	21.1
45.3	210.2	12.2	120.2
46.3	310.2	21.1	0.0
47.3	210.2	120.2	0.0
50.3	300.2	3.1	125.1
51.3	2.2	300.2	124.2
52.3	300.2	125.1	0.0
53.3	300.2	124.2	0.0
54.3	300.2	3.1	21.1
55.3	2.2	300.2	120.2
56.3	300.2	21.1	0.0
57.3	300.2	120.2	0.0
60.2	17.1	314.0	0.0
61.3	12.2	314.0	0.0
62.2	5.1	314.0	0.0
63.1	314.0	0.0	0.0
64.3	310.2	3.1	0.0
65.3	300.2	12.2	0.0
66.3	310.2	1.1	0.0
67.3	310.2	0.0	0.0
70.3	300.2	3.1	5.1
71.3	2.2	300.2	104.2
72.3	300.2	5.1	0.0
73.3	300.2	104.2	0.0
74.3	300.2	3.1	0.0
75.3	2.2	300.2	0.0
76.3	300.2	1.1	0.0
77.3	300.2	0.0	0.0
100.2	17.1	63.1	252.0
101.3	14.2	60.2	252.0
102.3	250.2	5.1	21.1
103.3	210.2	14.2	60.2
104.2	63.1	252.0	0.0
105.3	60.2	252.0	0.0
106.3	250.2	21.1	0.0
107.3	210.2	60.2	0.0
110.3	240.2	5.1	63.1
111.3	4.2	240.2	62.2
112.3	240.2	5.1	21.1
113.3	4.2	240.2	60.2
114.3	240.2	63.1	0.0
115.3	240.2	62.2	0.0
116.3	240.2	21.1	0.0
117.3	240.2	60.2	0.0
120.2	17.1	252.0	0.0
121.3	14.2	252.0	0.0
122.3	250.2	5.1	0.0
123.3	240.2	14.2	0.0
124.2	3.1	252.0	0.0
125.1	252.0	0.0	0.0
126.3	250.2	1.1	0.0
127.3	250.2	0.0	0.0
130.3	240.2	3.1	5.1
131.3	4.2	240.2	42.2
132.3	240.2	5.1	0.0
133.3	4.2	240.2	0.0
134.3	240.2	3.1	0.0
135.3	240.2	42.2	0.0
136.3	240.2	1.1	0.0
137.3	240.2	0.0	0.0
140.3	210.2	21.1	17.1
141.3	20.2	210.2	16.2
142.3	210.2	5.1	21.1
143.3	20.2	210.2	14.2
144.3	210.2	3.1	21.1
145.3	20.2	210.2	12.2
146.3	210.2	21.1	0.0
147.3	20.2	210.2	0.0
150.4	227.3	0.0	0.0
151.4	26.3	200.2	0.0
152.3	200.2	5.1	21.1
153.3	200.2	4.2	20.2
154.3	200.2	3.1	21.1
155.3	200.2	2.2	20.2
156.3	200.2	21.1	0.0
157.3	200.2	20.2	0.0
160.3	210.2	17.1	0.0
161.3	210.2	16.2	0.0
162.3	210.2	5.1	0.0
163.3	210.2	14.2	0.0
164.3	210.2	3.1	0.0
165.3	210.2	12.2	0.0
166.3	210.2	1.1	0.0
167.3	210.2	0.0	0.0
170.3	200.2	3.1	5.1
171.3	200.2	2.2	4.2
172.3	200.2	5.1	0.0
173.3	200.2	4.2	0.0
174.3	200.2	3.1	0.0
175.3	200.2	2.2	0.0
176.3	200.2	1.1	0.0
177.3	200.2	0.0	0.0
200.2	17.1	63.1	125.1
201.3	12.2	60.2	104.2
202.3	14.2	60.2	125.1
203.3	14.2	60.2	104.2
204.3	12.2	120.2	68.1
205.3	12.2	42.2	120.2
206.4	171.3	0.0	0.0
207.3	10.2	60.2	120.2
210.2	63.1	125.1	0.0
211.3	62.2	104.2	0.0
212.3	60.2	125.1	0.0
213.3	60.2	104.2	0.0
214.3	120.2	63.1	0.0
215.3	42.2	120.2	0.0
216.3	60.2	120.2	21.1
217.3	60.2	120.2	0.0
220.3	42.2	104.2	17.1
221.3	12.2	42.2	104.2
222.4	155.3	0.0	0.0
223.3	40.2	14.2	104.2
224.4	153.3	0.0	0.0
225.3	100.2	12.2	42.2
226.4	51.3	100.2	0.0
227.3	10.2	40.2	100.2
230.3	42.2	104.2	5.1
231.3	42.2	104.2	0.0
232.3	40.2	104.2	5.1
233.3	40.2	104.2	0.0
234.3	100.2	42.2	3.1
235.3	100.2	42.2	0.0
236.3	40.2	100.2	1.1
237.3	40.2	100.2	0.0
240.2	17.1	125.1	0.0
241.3	16.2	120.2	0.0
242.3	14.2	125.1	0.0
243.3	14.2	120.2	0.0
244.3	12.2	120.2	21.1
245.3	12.2	120.2	0.0
246.3	10.2	120.2	21.1
247.3	10.2	120.2	0.0
250.2	3.1	125.1	0.0
251.3	2.2	124.2	0.0
252.0	Input vle ariaXb_1		
253.3	124.2	0.0	0.0
254.3	120.2	3.1	0.0
255.3	2.2	120.2	0.0
256.3	120.2	21.1	0.0
257.3	120.2	0.0	0.0
260.3	104.2	17.1	0.0
261.3	12.2	104.2	0.0
262.3	14.2	104.2	5.1
263.3	14.2	104.2	0.0
264.3	100.2	12.2	3.1
265.3	100.2	12.2	0.0
266.3	10.2	100.2	1.1
267.3	10.2	100.2	0.0
270.3	104.2	3.1	0.0
271.3	2.2	104.2	0.0
272.3	104.2	5.1	0.0
273.3	104.2	0.0	0.0
274.3	100.2	3.1	0.0
275.3	2.2	100.2	0.0
276.3	100.2	1.1	0.0
277.3	100.2	0.0	0.0
300.2	17.1	63.1	0.0
301.3	16.2	60.2	0.0
302.3	14.2	60.2	21.1
303.3	14.2	60.2	0.0
304.3	12.2	63.1	0.0
305.3	12.2	60.2	0.0
306.3	10.2	60.2	21.1
307.3	10.2	60.2	0.0
310.2	5.1	63.1	0.0
311.3	4.2	62.2	0.0
312.3	60.2	5.1	0.0
313.3	4.2	60.2	0.0
314.0	Input variable X_2		
315.3	62.2	0.0	0.0
316.3	60.2	21.1	0.0
317.3	60.2	0.0	0.0
320.3	42.2	17.1	0.0
321.3	14.2	42.2	0.0
322.3	40.2	14.2	5.1
323.3	40.2	14.2	0.0
324.3	12.2	42.2	3.1
325.3	12.2	42.2	0.0
326.3	10.2	40.2	1.1
327.3	10.2	40.2	0.0
330.3	42.2	5.1	0.0
331.3	4.2	42.2	0.0
332.3	40.2	5.1	0.0
333.3	40.2	4.2	0.0
334.3	42.2	3.1	0.0
335.3	42.2	0.0	0.0
336.3	40.2	1.1	0.0
337.3	40.2	0.0	0.0
340.2	21.1	17.1	0.0
341.3	20.2	16.2	0.0
342.3	14.2	21.1	0.0
343.3	20.2	14.2	0.0
344.3	12.2	21.1	0.0
345.3	20.2	12.2	0.0
346.3	10.2	21.1	0.0
347.3	10.2	20.2	0.0
350.2	3.1	5.1	21.1
351.3	2.2	4.2	20.2
352.2	5.1	21.1	0.0
353.3	4.2	20.2	0.0
354.2	3.1	21.1	0.0
355.3	2.2	20.2	0.0
356.2	21.1	0.0	0.0
357.3	20.2	0.0	0.0
360.0	Input variable X_3		
361.3	16.2	0.0	0.0
362.3	14.2	5.1	0.0
363.3	14.2	0.0	0.0
364.3	12.2	3.1	0.0
365.3	12.2	0.0	0.0
366.3	10.2	1.1	0.0
367.3	10.2	0.0	0.0
370.2	3.1	5.1	0.0
371.3	2.2	4.2	0.0
372.2	5.1	0.0	0.0
373.3	4.2	0.0	0.0
374.2	3.1	0.0	0.0
375.3	2.2	0.0	0.0
376.2	1.1	0.0	0.0
377.0	Input ONE		

of interconnections among switching elements in such a way that the specified output signal, as defined by the logical function, is eventually obtained. If a catalog for logical functions has been generated to resemble previous routines, the wiring table can be directly derived by taking five steps:

1. Write the truth table of the desired output function versus input variables X_1, X_2, and X_3, similar to Table IV.
2. Split the binary function table into groups of three bits and convert into octal index z.
3. Derive level L directly from the catalog of Table III for index z.
4. List required input functions as z:L in Table IV—A.
5. Repeat steps 3 and 4 for each input until $L=0$ as shown in Table IV—A.

Shown in Fig. 1 are entries in a circuit diagram, feeding from output toward input. Level index L must decrease eventually from 3 toward 0, which designates independent variable X_1, X_2, and X_3, in addition to 0 and 1.

Don't Care Conditions

A logical function which is not completely defined for all possible input combinations is frequently characterized by the term DON'T CARE. If the catalog of logical functions is applicable for p input variables, the minimum solution is immediate from evaluation of all suitable functions Y as shown in Table V. Whenever the output is irrelevant for a certain input combination, the symbol 2 is inserted in the function table to designate the DON'T CARE condition, meaning that *either* 0 *or* 1 is acceptable as the logical output. Because two bits are not specified in this example, four output functions would be satisfactory.

Table V tabulates these four function matrices and their equivalent octal indexes z. Look-up in the catalog indicates that one of the four output functions is of level $L=1$; the rest are of level 3. Accordingly, the level $L=1$ function is the MINIMUM solution.

Microprogramming

Certain types of digital devices operate with a single logical element and an auxiliary memory. Instead of interconnecting several logical elements as in Fig. 1, the intermediate output functions may be stored in a memory and applied to a single logical element, three at a time, if the latter has three input terminals. Providing the cost for the memory is negligible and speed requirements can be met with this serial operation, certain advantages of microprogramming are apparent. The microprogram may be derived from the combinational network or directly from this catalog of logical functions. Then, for each intermediate logical function, a memory location has to be provided in the microprogrammed machine.

Conclusions

Use of logical design catalogs, as generated by computer routines, offers an expedient method for certain types of design problems. The obvious restriction is in the limited number of input variables. To handle function matrices of large size, certain algorithms have been derived which are an inversion of the expansion theorem previously developed.[2] This scheme will be described in a later paper on partitioning of complex-function matrices.

References

1. MATRIX ANALYSIS OF LOGICAL NETWORKS, E. J. Schubert. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Mar. 1958, pp. 10–13.
2. MATRIX SYNTHESIS OF HIGH-SPEED LOGIC, E. J. Schubert. *Ibid.*, vol. 78, Mar. 1959, pp. 4–8.
3. UNIVERSAL MODULUS, MATRIX LOGIC VI, E. J. Schubert. *Ibid.*, vol. 79, 1960 (Jan. 1961 section), pp. 778–80.

Generalized Recurrence Relations in the Analysis of Nonlinear Systems

ALFRED A. WOLF
MEMBER AIEE

INVOLVED IN SOLVING a certain class of nonlinear systems with nonlinear derivatives, among other things, are the general recurrence relations presented herewith. The theory is applied to systems involving circuits and feedback systems. Incidental to the paper, a relationship is developed between circuits and feedback systems.

In 1960,[1] the author developed a theory of nonlinear systems involving recurrence relations to solve a certain class of systems exactly, the nonlinearities being taken as polynomials as shown in equation 4. That theory was easily extended to systems involving nonlinearities, which included both derivative and functional polynomials of the response variable. A mathematical theory[2] was evolved earlier to analyze a broad class of nonlinear systems, and included a principle of stability and indications for a synthesis theory.

The recurrence equations developed here by partition can alternatively be solved by the Taylor-Cauchy transform.[3] This is comparable to the Laplace transform solution for linear continuous systems.[8,10] By a suitable transformation of variables, as shown in reference 4, the present theory also can be extended to nonlinear systems involving transcendents.

Theorem—System Restrictions

A class of systems, here considered, may be described by an ordinary nonlinear differential equation such as

$$Z(D)x(t)+F\{x,\dot{x},\ldots,x^{(m)}\}=g(t) \tag{1}$$

where

$$Z(D)=\sum_{n=0}^{N_1} f_nD^n \tag{2}$$

$$D=\frac{d}{dt};\ D^n=\frac{d^n}{dt^n} \tag{3}$$

where m denotes order of the derivative of the nonlinear part.

In reference 1, F was taken as a polynomial in x only, as given by

---

Paper 60-193, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted November 4, 1959; made available for printing February 21, 1961.

ALFRED A. WOLF, formerly with the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa., is with Emertron Inc., Emerson Research Laboratories Division, Silver Spring, Md

The author is grateful for support and help offered by Dr. Harry Sohon, Dr. Y. H. Ku, Dr. J. G. Brainerd, and J. Dietz, all of the University of Pennsylvania; Dr. T. J. Higgins of the University of Wisconsin; Dr. L. Kazda of the University of Michigan; and K. Lord, Dr. R. Weller, G. Peck, and L. Holmes of the Stromberg-Carlson Company; also to the National Science Foundation for its support.

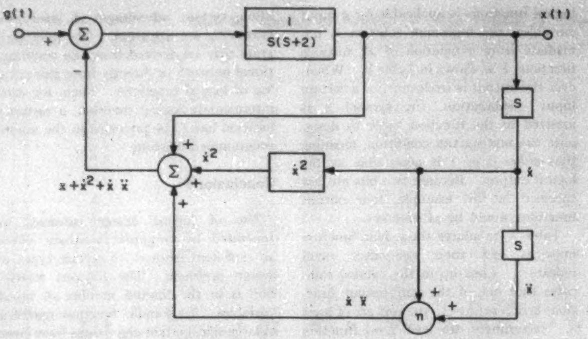


Fig. 1. A second-order servo with square-law and multiplicative feedback

$$F(x)=\sum_{k=1}^{N} a_k x^k \tag{4}$$

If the linear and nonlinear parts of equation 1 and its driving function satisfy certain conditions, then its solution is found as subsequently shown.

Moment Theorem

Given equation 1 suitably restricted and $y(t)$, the impulsive response of $Z(D)$, then this is solved by the linear combination of all the moments of the folded impulsive response of the linear part. Thus

$$x(t)=\sum_{n=0}^{\infty} c_n Q_n(t) \tag{5}$$

where

$$Q_n(t)=\int_0^t \tau^n y(t-\tau)d\tau \tag{6}$$

and

$$y(t)=\frac{1}{2\pi j}\int_{Br}\frac{e^{st}}{Z(s)}ds \tag{7}$$

where Br is the Wagner Bromwich contour.

The c_n's are given by equation 8 if the system is initially quiescent and if F is given by equation 4.

$$C_n=b_n-\sum_{k=0}^{N} c_{n-\alpha}{}^{(k)} a_k \tag{8}$$

where b_n's are the Taylor coefficients of the expansion of $g(t)$, the forcing function, and α is a positive integer or zero related to the order and degree of the system.

Restrictions on System

1. Operator $Z(D)$ is restricted in such a manner that the f_n's cause $y(t)$ to be of exponential type and of order one.
2. Forcing function $g(t)$ is of exponential type.
3. F is a single-valued, continuous nonlinear function of x and is required to satisfy the further condition that: Given a pair of positive constants $M>0$ and $a\geq 0$ and a pair of functions $u(t)$ and $v(t)$ which are continuous and asymptotically like $e^{-\beta t}$, then

$$|F\{u(t)\}-F\{v(t)\}|<Me^{-at}|u(t)-v(t)|$$

for all $t\geq 0$.

For the case where derivatives are involved in a nonlinear manner, F must be a single-valued, continuous function of x and of its derivatives.

Nonlinear Terms Containing Derivatives

Further aspects of the subject theory will be introduced. Specifically, how do nonlinear terms affect the method of solution when they contain derivatives of the dependent variable? A moderately difficult example—a second-order servomechanism with square-law and multiplicative feedback—is analyzed, illustrating use of the new formulas. An equivalent circuit, which has the same response to the given input and initial conditions, is derived

Two possible nonlinear functions, consistent with the conditions given in reference 2, are considered. For the first

$$F(x)=\sum_{m=0}^{M}\sum_{k=0}^{N} a_{mk}\{x^{(k)}\}^m \tag{9}$$

and, for the second,

$$F(x)=\sum_{m=0}^{M} a_m x^m \sum_{k=0}^{N} d_k x^{(k)} \tag{10}$$

where $x=x(t)$, and the superscript (k) indicates the kth derivative of x with respect to time, equation 10 may be written

$$F(x)=\sum_{m=0}^{M}\sum_{k=0}^{N} a_m d_k x^m x^{(k)} \tag{11}$$

The discussion to follow will make use of 9 or 10, with no change in the method given in reference 1. Substituting $F(x)$ in 9 into 1

$$Z(D)x(t)+\sum_{m=0}^{M}\sum_{k=0}^{N} a_{mk}\{x^{(k)}\}^m=g(t) \tag{12}$$

Equation 12 may be rewritten

$$Z(D)x(t)=g(t)-\sum_{m=0}^{M}\sum_{k=0}^{N} a_{mk}\{x^{(k)}\}^m \tag{13}$$

and the auxiliary function $A(t)$ becomes as shown in references 1 and 2

$$A(t)=\sum_{n=0}^{\infty} c_n t^n=g(t)-\sum_{m=0}^{M}\sum_{k=0}^{N} a_{mk}\{x^{(k)}\}^m \tag{14}$$

in which $g(t)$ is expandable to

$$g(t)=\sum_{n=0}^{\infty} b_n t^n \tag{15}$$

Under these conditions, the solution of 12 is given by

$$x(t)=\sum_{n=0}^{\infty} c_n Q_n(t) \tag{16}$$

$$c_n=b_n-\sum_{m=0}^{M}\sum_{k=0}^{N} a_{m,k-\alpha} c_{n,k}{}^{(m)} \tag{17}$$

where

$$Q_n(t)=\int_0^t \tau^n y(t-\tau)d\tau \tag{18}$$

The formulas for the coefficients $c_{n,k}{}^{(m)}$ are developed in reference 2 in terms of c_n.

In the derivation of equation 17, the initial conditions were assumed to be zero. If they are not zero, 17 is correspondingly altered. The case of nonzero initial conditions is discussed in detail in reference 2 and illustrated here in the example.

If equation 11 is used instead of 9, recurrence formulas are, instead of 17, given by

$$c_n=b_n-\sum_{m=0}^{M}\sum_{k-0}^{N} a_{m-\alpha} d_{k-\beta} h_{n,k}{}^{(m)} \tag{19}$$

where $\alpha,\beta>0$, and the h's are functions of the c's.

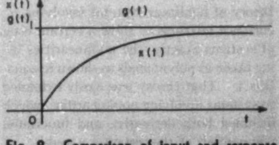


Fig. 2. Comparison of input and response functions

Second-Order Servomechanism

A nonlinear servomechanism is shown in Fig. 1. If the initial conditions are

$$x(0)=0$$

$$\dot{x}(0)=1 \qquad (20)$$

what is the response of the system to a unit-step function? (See Fig. 2.)

The forward loop of this system can be a motor which requires little amplification, and a double-tachometer can provide velocity and acceleration feedback. The velocity feedback is through a square-law device and, because of multiplication, a second nonlinear term results.

The differential equation governing the system dynamics is

$$\frac{d^2x}{dt^2}+2\frac{dx}{dt}+x+\left(\frac{dx}{dt}\right)^2+\left(\frac{dx}{dt}\right)\left(\frac{d^2x}{dt^2}\right)=u(t) \qquad (21)$$

and it was derived by partitioning the network and considering the parts separately.[2]

Following the operational notation used in 1, equation 21 becomes

$$D^2x+2\dot{x}+x+(\dot{x})^2+\dot{x}\ddot{x}=1;\ t>0 \qquad (22)$$

where the super dots are time derivatives in the partitioned part, and the *D*'s are used in the nonpartitioned part. The order of the differentiation corresponds to the number of dots.

Partitioning 22 to the left of the $2\dot{x}$ term results in

$$D^2x=1-(2\dot{x}+x+(\dot{x})^2+\dot{x}\ddot{x}) \qquad (23)$$

The auxiliary function then is

$$\sum_{n=0}^{\infty} c_n t_n=1-2\dot{x}-x-(\dot{x})^2-\dot{x}\ddot{x} \qquad (24)$$

The impulsive response of D^2x, the linear partitioned part, is

$$y(t)=\frac{1}{2\pi j}\int_{Br}\frac{e^{ts}}{s^2}\,ds \qquad (25)$$

from which

$$y(t)=t;\ t\geq 0 \qquad (26)$$

From 26, the moments are easily calculated. Thus

$$Q_n(t)=\int_0^t \tau^n y(t-\tau)d\tau \qquad (27)$$

$$Q_n(t)=\int_0^t \tau^n(t-\tau)d\tau \qquad (28)$$

$$Q_n(t)=t\int_0^t \tau^n d\tau-\int_0^t \tau^{n+1}d\tau \qquad 29)$$

$$Q_n(t)=\frac{t^{n+2}}{(n+1)(n+2)} \qquad (30)$$

Thus, the solution $x(t)$ is

$$x(t)=\sum_{n=0}^{\infty} c_n\frac{t^{n+2}}{(n+1)(n+2)}+K_0+K_1t \qquad (31)$$

Boundary conditions in equation 20, when applied to 31, make clear that

$$K_0=0$$

and that

$$K_1=1$$

Thus, on substituting these values, 21 becomes

$$x(t)=\sum_{n=0}^{\infty} c_n\frac{t^{n+2}}{(n+2)(n+1)}+t \qquad (32)$$

Recurrence Relations

From 22 come equations 33 and 34

$$\dot{x}(t)=\sum_{n=0}^{\infty} c_n\frac{t^{n+1}}{n+1}+1 \qquad (33)$$

$$\ddot{x}(t)=\sum_{n=0}^{\infty} c_n t^n \qquad (34)$$

Substituting 22, 23, and 24 into auxiliary equation 17 and noting that

$$\left[\sum_{n=0}^{\infty}\frac{c_n}{n+1}t^{n+1}\right]^2=\sum_{n=0}^{\infty} c_{n,1}{}^{(2)}t^{n+2} \qquad (35)$$

and that

$$\left[\sum_{n=0}^{\infty}\frac{c_n}{n+1}t^{n+1}\right]\left[\sum_{n=0}^{\infty} c_n t^n\right]$$

$$=\sum_{n=0}^{\infty} c_{n,1,2}t^{n+1} \qquad (36)$$

the process yields the recurrence relations.

For $n=0$

$$c_0=1-2-1-c_0 \qquad (37)$$

$$c_0=-1 \qquad (38)$$

For $n=1$

$$c_1=-2c_0-1-2c_0-c_{0,1,2}{}^{(2)}-c_1 \qquad (39)$$

$$c_1=1 \qquad (40)$$

For $n\geq 2$, the generalized recurrence formula,

$$c_n=-\frac{2c_{n-1}}{n}-\frac{c_{n-2}}{(n-1)n}-c_{n-2,1}{}^{(2)}-$$

$$\frac{2c_{n-1}}{n}-c_{n-1,1,2}{}^{(2)}-c_n \qquad (41)$$

Symbols $c_{n-2,1}{}^{(2)}$ and $c_{n-1,1,2}{}^{(2)}$, which are explained in reference 2, mean here

$$c_{n-2,1}{}^{(2)}=\sum_{k=0}^{n-2}\frac{c_k}{k+1}\,\frac{c_{n-k-2}}{n-k-1} \qquad (42)$$

$$c_{n-1,1,2}{}^{(2)}=\sum_{k=0}^{n-1}\frac{c_kc_{n-k}}{k+1} \qquad (43)$$

Calculating the next several coefficients $c_2, c_3, c_4, \ldots$, etc., yields

$$c_2=-\frac{2c_1}{2}-\frac{c_0}{2}-c_0{}^2-c_2-c_0c_1-\frac{c_1c_0}{2}=-\frac{1}{2}$$

$$c_3=-\frac{2c_2}{3}-\frac{c_1}{2\cdot 3}-\frac{c_0c_1}{2}-\frac{c_1c_0}{2.1}-\frac{2}{3}c_2-c_0c_2-$$

$$\frac{c_1{}^3}{2}-\frac{c_1c_0}{3}-c_3=\frac{1}{2\cdot 3}$$

Substituting these coefficients into

$$x(t)=t+\sum_{n=0}^{\infty}\frac{c_n}{(n+1)(n+2)}t^{n+2}$$

yields

$$x(t)=t-\frac{t^2}{2!}+\frac{t^3}{3!}-\frac{t^4}{4!}+\frac{t^5}{5!}-\ldots-\frac{(-t)_n}{n!}+\ldots$$

The *n*th term is obtained by induction, it being noted that

$$e^{-t}=1-t+\frac{t^2}{2!}-\ldots+(-1)^n\frac{t^n}{n!}+\ldots$$

from which it is recognized that

$$x(t)=1-e^{-t} \qquad (44)$$

This result is known to be correct since it identically satisfies the differential equation which governs the system behavior; see equation 21.

Method of Analysis

A number of questions arose in a paper previously mentioned.[1] One such question, which evolved into the subject of this paper, is concerned with nonlinear functions of derivatives of arbitrary order entering into a given physical system. These arise physically in many places in nature. One example is tachometer feedback through a nonlinear device, and another occurs in mechanics when there is drag of a body through a viscous fluid. The damping force, in such cases, is a nonlinear function of the velocity. Because of cases like these, it was necessary to show that the theory previously presented still held. Space limitations forbid rigorous treatment in this paper, but adequate material may be found in the author's doctoral dissertation.[3]

Restrictions and, hence, limitations of the analysis are discussed in the concluding section of this paper. The restrictions are sufficient but not necessary. No set of necessary conditions has been discovered so far. The restrictions given here limit the analysis to systems that are physically realizable. This is one reason

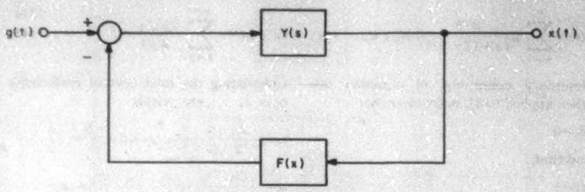


Fig. 3. Canonical feedback system

the method is important as a general technique for analyzing physical systems.

Analysis of the example shows the method to be straightforward in procedure and the results simply and quickly obtained. Such things as a full theory on initial conditions and their effect on the behavior of the solutions are not treated here. However, in the example, an indication of how initial conditions may be treated is indicated.

Circuits and Feedback Systems

As a by-product of this work, development of relations between circuit configurations and feedback system configurations proves interesting. The first step in this direction is to consider equation 1, which is composed of both a linear and nonlinear set of terms and a forcing function. Solution of the feedback system given in Fig. 3 gives rise to an integral equation that is the equivalent of equation 1. Similarly, solution of the circuit given in Fig. 4 gives rise to equation 1. Therefore, Figs. 3 and 4 are equivalent. This establishes a relation between circuits and feedback systems.[2] The output of the feedback system is equivalent to the current response of the circuit. Fig. 5 represents the circuit equivalent of Fig. 1.

Conclusions and Comments

Some aspects have been presented of general recurrence relations which occur in solving a certain class of nonlinear systems, the theory being a direct extension of work reported in reference 1. Again, the partition process leads to solvable recurrence relations. This stems from the fact that rearranging terms of the nonlinear differential equation to form the auxiliary or modified driving function does not include a shift of the highest-order derivative to the right member.

Since this derivative remains on the left member of the differential equation, the resulting relation, which is a nonlinear difference equation, has recursive properties. This means that the value of any coefficient in the recurrence equation depends for its value only on previous coefficients. With a knowledge of the boundary values—for example, the initial values—the recurrent relations are always recursively solvable.

Convolution sums arise from nonlinear terms in the differential equation. When a nonlinear term equal to the highest-order linear term occurs, the convolution sum yields a nonlinear term in c_0. As a result, the first value solved for, namely c_0, is ambiguous. This ambiguity can be resolved only by appealing to the physics of the problem. If the nonlinear term contains a derivative of higher order than the highest linear term, the resulting recurrence formula cannot be solved recursively.[2] To ameliorate this condition, a change of variables in the original equation is generally helpful.[2]

The conditions required of the linear, nonlinear, and forcing function terms of the differential equation evidently restrict application of the theory to a limited class of problems.[4] Fortunately, however, the modified Lipschitz condition (condition 3)[1,2] describes physical nonlinearities, and it requires that the nonlinear function rate of change be not too rapid. Thus, sharp breaks in nonlinearity would not satisfy the Lipschitz condition, but since they are not too likely to occur in nature, this theory has broad applications in physical systems. In the case of ideally rapid switching nonlinearities, the theory is still usable with omission of the discontinuous points.[1,2]

When the forcing function is random,[5] the method given here may be combined with the transform-ensemble method to solve nonlinear random processes. If the modified driving function is expanded into a power series, the recurrence equations are formally obtained by the Taylor-Cauchy transform.[3] Combining the latter with the transform-ensemble theorem leads to an elegant treatment of the nonlinear, random problem solution.[11]

If recurrence coefficients terminate, the solution is obtained in closed form. On the other hand, if the closed form solution of the recurrence equation is obtained, the closed form solution of the differential equation can be found under certain conditions.[1,2,5]

Finally, the form of solution is markedly dependent upon initial values. In linear systems, initial conditions have essentially no bearing on the form of solution.[6–8] Thus, behavior of a nonlinear system may show radically different responses before and after a switching operation in which some elements may retain charge.

References

1. Recurrence Relations in the Solution of a Certain Class of Nonlinear Systems, **Alfred A. Wolf.** *AIEE Transactions*, pt. I (*Communica-*

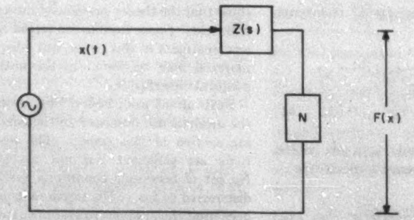


Fig. 4. Equivalent series circuit

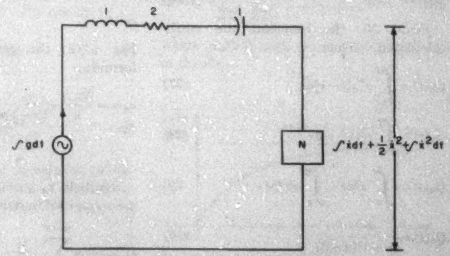


Fig. 5. Equivalent circuit of Fig. 1

tion and Electronics), vol. 78, 1959 (Jan. 1960 section), pp. 830–34.

2. A Mathematical Theory for the Analysis of a Class of Nonlinear Systems, A. A. Wolf. *Ph.D. Dissertation*, University of Pennsylvania, Philadelphia, Pa., June 1958.

3. Taylor-Cauchy Transforms for Analysis of a Class of Nonlinear Systems, Y. H. Ku, A. A. Wolf, J. H. Dietz. *National Convention Record, pt. 2*, (Circuit Theory), Institute of Radio Engineers, New York, N. Y., 1959, pp. 49–61; also *Proceedings*, vol. 48, May 1960, pp. 912–22.

4. Analysis of Transcendental Nonlinear Systems, Alfred A. Wolf. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, Nov. 1960, pp. 449–51.

5. Transform-Ensemble Method for the Analysis of Linear and Nonlinear Systems with Random Inputs, Y. H. Ku, A. A. Wolf. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 15, 1959, pp. 441–45.

6. On the Significance of Transients and Steady State Behavior in Nonlinear Systems, A. A. Wolf. *Proceedings*, Institute of Radio Engineers, vol. 47, no. 10, Oct. 1959, pp. 1785–86.

7. A Stability Criterion for Nonlinear Systems, Y. H. Ku, A. A. Wolf. *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 78, July 1959, pp. 144–48.

8. Some Recent Advances in the Analysis and Synthesis of Nonlinear Systems, Alfred A. Wolf. *Ibid.*, vol. 80, Nov., 1961.

9. Entire Functions (book), R. P. Boas. Academic Press, New York, N. Y., 1954.

10. Theorie und Anwendung der Laplace Transformation (book), G. Doetsch. Dover Publications, Inc., New York, N. Y., 1943.

11. Poles and Zeros of a Random Process in Linear and Nonlinear Systems, A. A. Wolf. *Proceedings*, National Electronics Conference, vol. 16, 1960, pp. 268–78.

Tentative Recommendation for the Definition and Testing of the Dynamic Performance of Analog-to-Digital Converters

WARREN M. GAINES
MEMBER AIEE

PETER P. FISCHER
MEMBER AIEE

ONE of the major objectives of the Component Specification Subcommittee of the Feedback Control Systems Committee is to promote improved definitions of the dynamic performance characteristics of the components which are useful to the engineer. In addition, the subcommittee assists in establishing guides which will yield more uniformity to the measurement and publication of such data. In accordance with this objective this report has been prepared on analog-to-digital converters.

General

During the last two decades the digital computer has emerged from the experimental to the fully operational stage. No longer used solely in "off-line" business, banking, or data processing applications, these computers now serve as an integral part of process control. Digital process-control computers, along with digital data-recorders, have found application in steel, chemical, utilities, and other industrial processes as well as in military systems. The language of these digital equipments is a discrete code while the processes that are controlled are continuous and have been traditionally instrumented with analog-type sensors. There exists, therefore, a need for devices that convert the continuous analog signals to discrete codes compatible with the digital equipment. With the exception of a few sensors which generate a discrete output directly, there are two major categories of devices for converting a continuous signal to a discrete code: (1) devices which convert shaft rotation to a digital code; and (2) devices which convert voltage to a digital code.

1.1 Scope

It is the intent of this report to suggest methods of classifying, specifying, and testing shaft-to-digital and voltage-to-digital converters.

1.2 Analog-to-Digital Conversion (ADC)

The term "digital" as used in this paper describes any system of quantizing information in discrete steps or bits. Analog-to-digital conversion refers to the process of quantizing a continuous function, thus forming a code consisting of discrete steps.

1.3 Types of Converters

There is a wide range of application of converters and, therefore, there exists a wide variety of converter types. For the purpose of this paper the two major classifications are considered separately; the voltage-to-digital converters and the shaft-position-to-digital converters.

1.3.1 Voltage-to-Digital Converters

These are various devices that convert an input analog voltage to a digitally coded signal.

1.3.2 Shaft-Position-to-Digital Converters

These are electromechanical devices which convert shaft rotation into a digitally coded signal. They have one or more code plates coupled directly or by means of gearing.

1.3.3 Types Not Included

A third class is the hybrid voltage-to-digital converter which uses a combination servomechanism and shaft encoder to accomplish the conversion. These devices encompass the problems of both the shaft encoder and the servomechanism. Normally, the servomechanism predominates the dynamic performance and, therefore, the definition of the dynamic performance of this type of device has been deferred until further work by the committee on the definition of the dynamic performance of servomechanisms. Pulse generators used for speed measurements or position location by external counters also have not been included because of their dependence on a large amount of external equipment and because of their special nature at the present state of the art.

1.4 Limitations

This paper contains a compilation, interpretation, and specification of the major dynamic and static characteristics of analog-to-digital converters. This document is not an equipment specification, and therefore does not include complete definition of environmental specification and physical packaging specifications.

Paper 61-15, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted September 6, 1960; made available for printing November 25, 1960.

W. M. Gaines is with the General Electric Company, Phoenix, Ariz.; P. P. Fischer is with the Westinghouse Electric Corporation, Baltimore, Md.

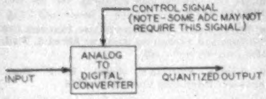

CONTROL SIGNAL
(NOTE - SOME ADC MAY NOT REQUIRE THIS SIGNAL.)
ANALOG TO DIGITAL CONVERTER
INPUT
QUANTIZED OUTPUT

Fig. 1. General analog-to-digital converter

Definitions

The block diagram of Fig. 1 typifies the general analog-to-digital converter. The continuous input signal is converted to a quantized output signal by the action of the converter. In order to initiate and control this action, control and/or enabling signals may be transmitted to the converter. These control signals may be electrically or manually initiated.

Fig. 2 is a typical analog-to-digital converter configuration for an electronic analog-to-digital converter. See the references for descriptions of several types of analog-to-digital converters.

2.1 TERMINOLOGY

The following definitions are presented to clarify the rather confusing terminology used to specify various aspects of system performance. Some of the definitions were derived from an unpublished report of Norden Laboratories.

Some terms refer only to shaft-position-to-digital converters (SDC), some only to voltage-to-digital converters (VDC), others are applicable to both. Applicability is indicated for each definition.

2.1.1 Total Count (SDC, VDC)

The total count is the full-scale reading of the converter. It is equal to the maximum number of discrete quantization increments available from the converter. For a binary coded ADC of n bits, the total count is 2^n.

2.1.2 Count per Turn (SDC)

The count per turn is the total number of discrete quantization increments produced by one revolution of the input shaft. For single-turn SDC, this is the same as the total count. For multiturn SDC, this is the total count divided by the number of turns.

2.1.3 Resolution (SDC, VDC)

The resolution is a measure of the smallest input change which can be detected, not necessarily measured, at the output of the ADC; see Fig. 3.

For VDC, it is a function of the signal-to-noise ratio, gain of the system, and the size of the least significant bit. It is expressed in terms of input voltage.

For SDC, it is a function of the size of the least significant bit. It is expressed as the reciprocal of the total count or as its angular equivalent.

2.1.4 Zero Offset (VDC)

The zero offset is the magnitude of the reading provided by the system, referred to the input, when the input is zero. It is a measure of the bias generated inside the measuring system, and is normally given in absolute units.

This definition has the advantage of reducing the zero offset of the measurement to an equivalent input error signal, irrespective of its actual origin in the system. Consequently, the relative importance of this type of inaccuracy is readily evaluated in comparison with the information signal itself. Furthermore, the preceding definition is compatible with that of the noise figure in amplifiers.

2.1.5 Positive Input Direction (SDC)

The positive input direction is the direction of motion of the input which produces an increasing count.

2.1.6 Zero Cross-Over Angle (SDC)

The zero cross-over angle is the position of input at which the output count changes to zero (for positive input direction). In a binary coded SDC, this is the point where the count changes from all "ones" to all "zeroes" (for positive input direction).

2.1.7 Cross-Over Angle (SDC)

The angular position of the input shaft (measured in a positive direction from the zero cross-over angle) at which a change in count should occur. Each quantization increment is associated with a nominal cross-over angle.

2.1.8 Value of Count (SDC, VDC)

The value of count is the nominal input value (either voltage or angle) which corresponds to a particular count. It is defined as the mid-point of the theoretical input region over which the output value occurs.

For an SDC this will be the arithmetic mean of the theoretical cross-over angles bracketing the region over which the count should occur.

For a VDC, it is the arithmetic mean of the theoretical input voltages bracketing the region over which the count should occur.

2.1.9 Linearity (VDC)

Linearity is a measure of the constancy of the ratio between output reading and input value. Based upon a straight line between the zero and total count (i.e., correcting for zero offset and gain errors), the linearity is defined as the deviation from this straight line expressed normally as a percentage of total count. This is the independent linearity as opposed to absolute or zero-based linearity which is obtained by using an uncorrected (for zero offset and gain errors) output.

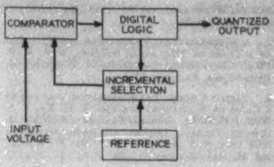


Fig. 2. Typical voltage-to-digital (VDC) configuration

2.1.10 Drift (VDC)

Drift is a measure of the change in zero offset as a function of time. It is normally given in absolute units.

2.1.11 Repeatability (VDC)

Repeatability or precision is a measure of reproducibility of several independent quantizations (counts) of a fixed input under fixed conditions. It is affected by such factors as friction and various types of hysteresis and nonrepetitive noise effects, as well as the digital resolution of the output.

Time duration as well as the other environmental and operating conditions must be specified in defining the repeatability. Repeatability (deviation between quantization) can be defined in absolute units or as a percentage of total count.

Note that a high degree of repeatability does not necessarily imply accuracy since an erroneous quantization could be highly repeatable.

2.1.12 Tolerance (SDC, VDC)

Tolerance or relative accuracy is the fractional amount, relative to total count, by which a quantization deviates from the true value of the input quantity after the zero offset has been corrected. The basis of comparison is the total count of the instrument, not the magnitude of the input quantity; see Fig. 4.

2.1.13 Deviation (SDC, VDC)

Deviation (per cent of reading; point accuracy) is the fractional amount, relative to the correct value of the signal at the point of measurement, by which the

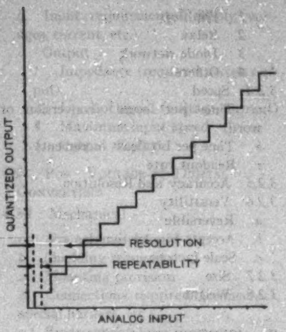


Fig. 3. Ideal analog-to-digital converter characteristic

measured value varies from the true value after zero offset has been corrected.

Fig. 4 indicates the difference between tolerance and deviation. Deviation is obviously more demanding on the equipment than the conventional tolerance since maximum error must decrease as the signal amplitude decreases.

2.1.14 Monotonicity (*SDC, VDC*)

Monotonicity is a measure of the ratio between output change and input change. It is an incremental linearity; i.e., the slope of the input-output relation. Since the output is incremental, it can be defined either as the slope of an equivalent linear curve drawn through the incremental function or as the variation in the ratio between the change in input required to generate the prescribed incremental change in output.

2.1.15 Quantization Uncertainty (*VDC, SDC*)

Quantization uncertainty is peculiar to any system which provides a discrete measure of a continuous variable. The discrete output levels are separated by regions of uncertainty of the input quantity. Those regions can be reduced by increasing the resolution of the system but they cannot be eliminated. When the input variable happens to fall in a region of uncertainty, the system output may indicate one or the other of two adjacent digital values. Ideally, an analog-to-digital converter should have the characteristic shown in Fig. 3. If all the steps are present and have the same height and the same width, the converter is linear and monotonic. Furthermore, if the separation between successive steps has zero width, a unique digital value corresponds to any value of the analog input. In practice, this is impossible, and conditions depicted in Fig. 5 occur.

Quantization uncertainty can be defined either as a fraction of the quantum or as the equivalent input signal. The latter is advantageous since it is in compatible form with the other measure of accuracy noted previously.

2.1.16 Maximum Allowable Input Rate (*SDC, VDC*)

The maximum allowable input rate is the change in input quantity in a unit time.

For SDC, this is specified in terms of rpm and depends upon the response of the readout methods used (bounce, natural frequency, etc., for brush-type readouts; photocell response time, readout circuit sensitivity, etc., for other types).

For VDC, this is specified in volts per unit time. It is dependent upon the response of the conversion circuits, etc.

2.1.17 Conversion Time (*VDC*)

The time to convert an analog signal to a digital code. Specify in unit time per bit or as unit time per conversion. Both are significant since some converters use a varied rate in order to allow longer settling times for the more significant bits.

2.1.18 Conversion Rate (*Readout Rate*) (*SDC, VDC*)

This is expressed in the number of discrete outputs, corresponding to discrete inputs, which may be obtained in a unit time.

For a VDC, this is the sum of input settling time plus total conversion time. In specifying conversion rate, care must be taken to establish whether special logical functions such as automatic range selection and polarity selection are included.

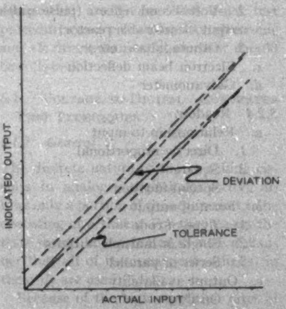


Fig. 4. Tolerance and deviation

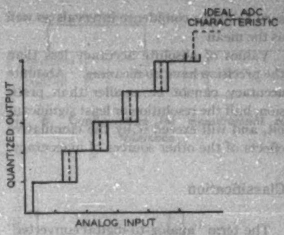


Fig. 5. Quantization uncertainty

Shaded area indicates the zone of uncertainty

2.1.19 Code (*SDC, VDC*)

Code is the relationship between a set of quantized signals and the value (of points) of the continuous input they represent.

2.1.20 Analog-to-Digital Conversion (*SDC, VDC*)

Analog-to-digital conversion is the process of quantizing (coding) a continuous function.

2.1.21 Total Converter Error (*SDC, VDC*)

The total converter error or accuracy is the maximum difference between the nominal (theoretical) input value which yields a given output (count) and the actual value (least or greatest) which will produce the same output. This may be expressed in input units or as a percentage of full-scale input.

2.1.22 Absolute Accuracy (*SDC, VDC*)

Absolute accuracy, or more correctly, absolute error is a measure of exactly how the quantized output of the system indicates the numerical value of the input, with reference to an accepted standard. The measure is expressed in the units of the full scale input. It is the absolute value of the total converter error. The absolute accuracy in units of the input is not a measure of quality, when considered alone (i.e., a voltmeter with an absolute accuracy of 0.1 volt is excellent if its full scale is 100 volts; it is mediocre if its full scale is 1 volt).

The absolute accuracy of equipment is determined by the input of the equipment, not the input to the transducer. To be unambiguous, the operation, conditions, and ranges under which the accuracy is achieved must be specified.

The absolute accuracy is the combination of all the contributing sources of inaccuracy in the equipment. The accuracy will normally not be a simple algebraic summation of the individual factors. The accuracy may therefore be statistical in nature, in which case it will be necessary to define confidence intervals as well as the mean.

Values of absolute accuracy less than the precision have no meaning. Absolute accuracy cannot be smaller than precision, half the resolution or least significant bit, and will exceed it by the cumulative effects of the other sources of inaccuracy.

Classification

The term "analog-to-digital converter" is too general to describe any specific piece of equipment. The following lists of classifications are presented as an attempt to array the many variations of ADC's in specific categories. Note that two main categories are used (VDC and SDC) and are followed by subcategories.

3.1 Shaft-Position-to-Digital Converters (SDC)

3.1.1 Number of increments (discrete positions in total count)

3.1.2 Code

a. Straight binary
b. Cyclic binary
c. Coded waveshapes
d. Special patterns (such as *V* drums to eliminate ambiguity)
e. Others (that is, octal, binary decimal, etc.)

3.1.3 Coding technique

a. Code wheel (rotational motion)
b. Code drum (rotational motion)
c. Code plate (linear motion)
d. Pulse wheel (working in conjunction with counter)
e. Cams
f. Phase shift (utilizing synchros, resolvers, etc.)

3.1.4 Type of Readout

a. Brushes
  1. Single brushes
  2. Multiple brushes
    i. Duplication for reliability
    ii. Special arrays to eliminate ambiguity such as *V* scan, etc.
b. Photoelectric (also photoresistive devices)
  1. Pulsed light
  2. Pulsed cell
  3. Continuous light and cell
c. Magnetic
d. Inductive
e. Counter

3.1.5 Type of Output

a. Relationship to input
  1. Directly proportional
  2. Sine or cosine
  3. Special relationships
b. Form of output
  1. Type of code
  2. Single or multiple functions (both sine and cosine)
  3. Series or parallel
c. Output availability
  1. On fly
  2. Under static conditions
d. Output compatibility
  1. Relays
  2. Flip-flops
    i. Vacuum tube
    ii. Transistors and other solid-state devices
  3. Magnetic tape input
  4. Counter
    i. Mechanical
    ii. Electrical (relay or electronic)

3.1.6 Accuracy and Resolution
3.1.7 Size
3.1.8 Weight
3.1.9 Readout rate

3.2 Voltage-to-Digital Converters (VDC)

3.2.1 Number of increments (discrete values in total count)

3.2.2 Code

a. Binary
b. Modified binary (cyclic, gray, etc.)
c. Coded waveshapes (pulse width, etc.)
d. Decimal or binary coded decimal
e. Others

3.2.3 Conversion Technique

a. Comparison
  1. Quantity
    i. Voltage
    ii. Current
  2. Method
    i. Incremental
    ii. "Half split"
    iii. Null balance
b. Time
  1. Sweep (Phantastron, capacitor discharge, etc.)
  2. Volt-second effect (pulse-width output of saturable reactor)
  3. Others (phase, etc.)
c. Electron beam deflection
d. Galvanometer

3.2.4 Readout

a. Relationship to input
  1. Directly proportional
  2. Square
  3. Special function
b. Form of output
  1. Type of code
  2. Single or multiple outputs
  3. Series or parallel
c. Output availability
  1. On fly
  2. With input static
d. Output device
  1. Flip-flop
  2. Relay
  3. Diode network
  4. Others

3.2.4 Speed

a. Time per count (conversion or word)
b. Time per bit (least increment)
c. Readout rate

3.2.5 Accuracy and Resolution

3.2.6 Versatility

a. Reversible
b. Accepts bipolar signals
c. Scale factors

3.2.7 Size
3.2.8 Weight

Specification

The control system designer requires various static and dynamic characteristics of components to be used in a design. The following are suggested groups of information which would constitute a minimum data sheet.

The following specifications should indicate the test conditions, test equipment, and circuits used, number of tests conducted, and method used to calculate specifications of statistical nature.

4.1 For Shaft-to-Digital Converter

4.1.1 Mechanical

a. Effective inertia at the input shaft
b. Starting and running torque required
c. Maximum slow speed (dependent only on mechanical considerations)
d. Concentricity of mounting diameters to shaft, and mounting provisions
e. Coupling provisions and requirements
f. Life
g. Environmental
h. Size and weight
i. Connections required (connector, wiring harness, etc.)

4.1.2 Functional

a. Nature of code
b. Parallel or serial readout
c. Total count
d. Scale factor (counts per revolution)
e. Output availability
  1. Sampled
  2. Continuous
f. Quality
  1. Resolution
  2. Absolute accuracy
g. Shaft direction for positive count

4.1.3 Operational

a. Peripheral equipment required (i.e., *V* scan brush selection logic)

b. Input requirements (supply voltages, current, etc.)
c. Output
  1. Impedance (required across output)
  2. Level of signal (voltage, current)
  3. Maximum input speed
  4. Output display

4.2 FOR VOLTAGE-TO-DIGITAL CONVERTERS

4.2.1 Mechanical

a. Size and weight
b. Shielding required
c. Mounting provision
d. Connections required (connectors, special harness, etc.)
e. Environmental requirements
f. Life

4.2.2 Functional

a. Nature of code
b. Parallel or series readout
c. Total count
d. Scale factor (volts per increment)
e. Output availability
  1. Sampled
  2. Continuous

f. Conversion time.

This is the maximum time and average time for one complete conversion. Careful consideration must be given the selection of the conversion times.

Significant system errors can be introduced by the sampling rate if it is lower than needed. Conversely, higher than necessary conversion rates can increase the cost of the equipment.

Ranging and polarity times if applicable should also be specified.

g. Quality.

Where it is necessary to meet performance requirements, absolute accuracy, tolerance or deviation, linearity, zero offset, monotonicity, repeatability, etc., should be specified within the following context.

Quality can be defined by either deviation or absolute accuracy or as the magnitudes of the individual errors which constitute the over-all "accuracies." The calculation of the total converter error will normally not be a simple algebraic summation of the individual errors and the exact method by which absolute accuracy can be determined should be indicated.

The quality should be specified for static as well as for dynamic conditions. "Static" in this case implies that the transients have been allowed to die down before a measurement of "quality" is made at a particular operating point. "Dynamic" implies operation of the equipment at maximum conversion rates and rates of change of the inputs. Measurements should be made over the operating range of the equipment. This should include not only the signal range but also different levels of resolution (sensitivity) if adjustment of resolution is available in the analog-to-digital converter, as well as variation with time and environmental conditions.

4.2.3 Operational

a. Peripheral equipment required (shift registers, logic, etc.)
b. Voltage level
  1. Input and output voltage level must be specified
  2. Output pulse shape and/or rise time where important
  3. Define all bit levels (0 and 1)
  4. Capabilities of handling both positive and negative input signals (bipolar)
c. Impedance
  1. Input and output impedance
  2. Dynamic or pulse impedance. Specify (where necessary for satisfactory application) source impedance including both upper and lower limits on source impedance
d. Control
  1. Type of control signal—read, convert, reset, etc.
  2. Signal characteristics—voltage level, rise time, fall time
  3. Program sequence

Testing

There are a variety of ways to measure the characteristics of an analog-to-digital converter, and thus the test technique used is necessarily an integral part of the specification. A number of suggested test techniques (of a general nature and not tied to a specific manufacturer's test equipment) are given. The usual test procedures and precautions for electronic and electromechanical equipment should be followed.

5.1 VOLTAGE-TO-DIGITAL CONVERTER TEST TECHNIQUES

5.1.1 General

The discrete nature of the switching circuits in analog-to-digital converters allows only a portion of the equipment to be operating at a time. For this reason the tests should be conducted in such a manner that all of the increments (digits) in the code are used.

Because of the high repetition rates of the clock on electronic ADC, extremely broad band equipment is needed to check waveshape and to investigate for "hidden" oscillations. Care must be exercised in the selection of test equipment to assure that there are no transients or other extraneous signals introduced. Particular care must be taken with the grounding of the signal sources.

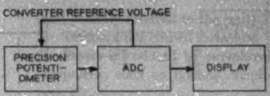


Fig. 6. Configuration for measurement of tolerance

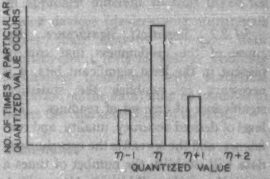


Fig. 7. Exaggerated histogram of large number of conversions with a fixed analog input signal showing the statistical spread due to quantization uncertainty

5.1.2 Static Test

5.1.2.1 Instrumentation. Absolute accuracy measurements require use of compatibly accurate voltage measuring instruments and secondary standards. Instruments 10 to 1 or better in absolute accuracy are desirable. For the most accurate analog-to-digital converters, this would require instrumentation with ±0.005- to 0.001-per-cent absolute accuracy. The latter figure is on the borderline of practicality with primary standards.

In measuring tolerance, the internal reference of the analog-to-digital converter can be used. In this case the input signals can be derived by a resistance bridge, see Fig. 6.

The bridge must be calibrated to ±0.005 per cent or better for checking the more accurate analog-to-digital converters.

The voltage source for the input should be selected for stability and freedom from ripple and noise. The ripple and noise peak-to-peak value should be less than 10 to 1 below the equivalent magnitude of the least significant bit of the converter.

The use of a printer or punch to provide a permanent record of the output is valuable if an extensive series of tests is to be made.

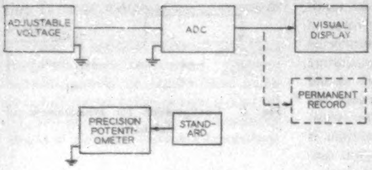


Fig. 8 (left). Configuration for measurement of absolute accuracy

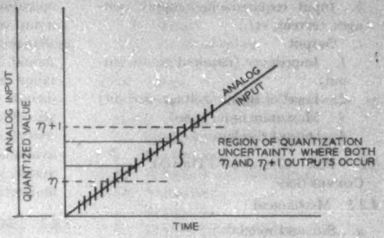


Fig. 9 (right). Interpretation of ramp test

Slashes indicate conversions

5.1.2.2 Statistical Significance. Because of the randomness that can be present in the least significant bits, it is necessary to establish the statistical significance of any set of readings. The level of desired accuracy, quality, and confidence, and the spread in the experimental data will establish the number of times a particular point will have to be repeated. A typical, although exaggerated, histogram of the output of an analog-to-digital converter with a constant input is shown in Fig. 7. It is assumed in the following that a sufficient number of tests are made to establish the desired confidence. It may be necessary to define the accuracy as a mean value plus a measure of the spread about the mean (e.g., standard deviation).

5.1.2.3 Static Quantization Uncertainty. By slowly increasing (decreasing) the input voltage, determine if all possible valid coded output values can be obtained (i.e., that there are no jumps) and that there are no reversals in the digital output.

Quantization uncertainty can be specified either as the change in input signal required to achieve an unambiguous change in the digital value or, in cases where there is a wider range of statistical uncertainty, by specifying the ratio of number of times the mid-value appears to the number of times adjacent values occur; see Fig. 7.

5.1.2.4 Static Monotonicity. Measure input voltage versus digital output value. Since the output is discrete there can be several values of input that will give the same value of digital output. There are several methods used to circumvent this problem. The input can be adjusted to approximately the mid-point by observing when the digital value changes, or the output of the comparator (or balance detector) can be observed to determine the point of minimum offset.

The incremental difference between successive readings then defines the static monotomicity. If there is a wide band of uncertainty or ambiguity, the adjustment can be made for the nth digital value by balancing the $(n + 1)$ and $(n = 1)$ values which occur over m conversions; see Fig. 7. The input signal that gives this balanced output can be used to calculate the slope of the input-output relation. A pseudo-wide band of uncertainty can be achieved by introducing dither into the input signal.

5.1.2.5 Static Linearity. The data obtained from *5.1.2.4* can be used directly to determine the static linearity.

5.1.3 Dynamic Testing

5.1.3.1 Instrumentation. See Fig. 8. The test of the converter under dynamic conditions requires the use of some means of recording the digital output at the testing rates. This may require magnetic tape or other high-speed recording media for the faster converters while paper tape or a printer may be satisfactory for the slower speed converters. The signal source should be capable of ramp and step changes and capable of generating timing signals. The signal source must be free of switching transients in the areas of interest and capable of being calibrated dynamically. The difficulties of achieving this objective are such that it is often more convenient to test the multiplexer and converter as a unit.

5.1.3.2 Measurement of Linearity and Monotonicity. The use of a ramp input voltage to measure linearity and monotonicity is suggested. The ramp should have a rate sufficiently low to allow several conversions having the same digital value to be made at each digital increment. The ramp generator may be electronic in the case of high-speed converters or mechanical for low-speed converters. The region of uncertainty, quantization uncertainty, can be estimated by observing the number of conversions in which there is reversal between (n) and $(n + 1)$. Assuming reasonable bands of uncertainty the monotonicity can be measured by the ratio of the number of conversions for each increment of digital value. If the band of uncertainty is large, it may be necessary to use the ratio of conversions for the next higher significant bit in order to obtain unambiguous results for the monotonic slope; see Fig. 9.

5.2 SHAFT-POSITION-TO-DIGITAL CONVERTER TEST TECHNIQUES

5.2.1 General

Shaft-position-to-digital transducers produce a large number of discrete outputs for one revolution. A 13-bit (binary coded) transducer can generate 8,192 discrete positions in each revolution. This approximately corresponds to dividing one revolution into 2-minute steps. Since the generation of each point, theoretically, is independent of the others, a complete check of positional accuracy for a 13-bit encoder would require a positioning device capable of producing 8,192 equal steps (per revolution) with an accuracy of not worse than 12 seconds of arc, and with ±1-minute adjustment at each step (adjustment readout to be compatible with the accuracy of the main step).

Testing each discrete position would involve a very tedious procedure as well as special precise positioning equipment of a quality beyond commercial availability. The following assumptions are made to facilitate a reasonable test procedure:

1. The code plate is made from a "master" which is absolutely accurate.
2. If distortion of the pattern is present on the code plate, it will appear at more than one of the discrete positions.
3. Errors other than those due to code-plate inaccuracy are due to readout head (brushes, photo cells, etc.) misalignment with respect to the code plate.

The following techniques will result in data from which the resolution, precision, absolute accuracy, etc., can be obtained.

5.2.1.1 Rotate the input shaft of the encoder at a constant speed, preferably the maximum allowable input speed. The transducer is now operated (supply voltages, etc., as specified) and the output of one track (bit) displayed on an oscilloscope. The output should appear as a square wave of constant frequency and amplitude. Variations in frequency indicate distortions in either the code pattern or the readout device alignment. Variations in amplitude may indicate scratches or dirt on the pattern. It is also conceivable that dirt or large scratches may cause frequency variation, although it is almost certain that this effect would be accompanied by variations in amplitude. This procedure, if repeated for each track, will give an indication of the functional accuracy of the transducer. It is assumed that distortion or misalignment between tracks is not possible if each track is free from distortion.

5.2.1.2 A test similar to that described in the preceding paragraph could be used to check relative alignment of each track. The pattern of each track would be compared in time phase to a track chosen as the reference track. (For test purposes, the least significant track might prove best suited unless the transducer has a built-in reference track or readout device.) The phase error thus measured is indicative of the transducer error.

5.2.1.3 It is conceivable that it may be desirable to test the instantaneous accuracy of an ADC. This may be done by rotating the input shaft at some constant speed and recording the output of each channel by means of a high-speed oscillograph. (The speed of rotation should be such that the frequency generated by the least significant bit track is well within the bandwidth of the oscillograph.) The recorded results can be analyzed and phase and amplitude distortions converted into ADC errors.

5.2.1.4 A combination of one of the preceding tests and a static discrete position check (at several selected points) would serve as an absolute functional check of ADC error.

5.2.1.5 Another test which is quite simple and effective is to display the output of the ADC on some device such as a bank of neon indicators, sensitive relays, etc. All the discrete counts (in the proper code) should appear as the ADC is rotated slowly (possibly by hand) from its zero position in a positive direction. The unit may be considered operating properly if there are no discontinuities in count larger than two resolution units.

Paragraphs *5.2.1.1* through *5.2.1.5* indicate several means of dynamically checking an ADC both for functional error and for noise. These errors may be due to some of the following causes:

a. Response of the readout device
  1. Brush
  2. Photoelectric device
  3. Magnetic head
b. Dirt or scratches between the code plate and the readout device
c. Variations in supply parameters
  1. Voltage variation
  2. Light source variation (due to age, supply voltage)
  3. Magnetic variations (stray field, demagnetization, etc.)
  4. Faulty connection to the unit (and possibly internal to the unit)
  5. Temperature-caused variations
  6. Errors introduced by peripheral equipment, such as readout amplifiers, etc.

5.3 ERROR TESTING

The testing of the dynamic and static functional error present in ADC requires an approach different to that taken for analog devices (devices with continuous rather than discrete outputs). Both point-by-point tests as well as sampling techniques (combined with statistical operation on the data) may be required to define the characteristics of an ADC.

Comments

The authors wish to underline the fact that the procedures, etc., outlined in this paper are *suggested* by the authors. It is not their purpose to attempt to write a "standard" for ADC, but rather to present the approach to testing and specifying ADC which is compatible with the viewpoint of the system designer and the component manufacturer. It is hoped that this paper will stimulate interest in initiating work on a "standard." The authors would appreciate receiving comments to assist in further work toward this end.

References

1. NOTES ON ANALOG-TO-DIGITAL CONVERSION TECHNIQUES (book), Alfred K. Susskind. Technology Press, Cambridge, Mass.

2. HANDBOOK OF AUTOMATION, COMPUTATION, AND CONTROLS, VOL. 2 (book), edited by E. M. Grabbe, S. Ramo, D. E. Wooldridge. John Wiley & Sons, Inc., New York, N. Y.,1959.

◆

Discussion

Charles A. Walton (International Business Machines Corporation, San Jose, Calif.): In spite of the great deal of discussion of the "accuracy" and performance of analog-to-digital converters, there is no pressing justification for discarding the definitions of accuracy, drift, and resolution which are well established for conventional instruments. The analog-to-digital converter should be treated as an instrument or, in the broadest sense, as a system which is equivalent to an instrument-plus-an-encoding-observer.

The accuracy definition from the "Electrical Engineers Handbook"[1] is:

"The majority of American manufacturers favor the simple definition of accuracy as the percentage ratio of the limit of error at any point on the scale to the full-scale value.... In instruments having a uniform scale this is equivalent to saying that no division mark is out of its proper position by more than a specified distance."

Also from the Handbook:[1]

"The 'error of indication' is found by subtracting from the value which it indicates the true value of the quantity measured by it, as determined by reference to some standard of a higher order than the instrument itself."

The drift definition from the "Process Instruments and Controls Handbook"[2] is: "Drift: The deviation of the instrument indication with time from an initial value when the measured variable and ambient conditions are constant."

Resolution is not ordinarily defined for an instrument. Webster's[3] definition of resolution, for optics, is: "The act or property of rendering visible the separate parts of an object...."

Repeatability is not ordinarily defined for instruments. It must be less than the accuracy, of course, and is similar to drift. Over a long period it is equal to the drift figure; over a short period it is less than the drift figure. Its value is quite important to process control work because derivatives, or slopes, with respect to time or otherwise, are often more important than the absolute value.

REFERENCES

1. ELECTRICAL ENGINEERS' HANDBOOK. ELECTRIC POWER (book), Harold Pender, William A. Del Mar, editors. John Wiley & Sons, Inc., New York, N. Y., pp. 5-26.

2. PROCESS INSTRUMENTS AND CONTROLS HANDBOOK (book), D. M. Considine, editor. McGraw-Hill Book Company, Inc., New York, N. Y., 1957.

3. WEBSTER'S NEW INTERNATIONAL DICTIONARY (second edition, unabridged). G. & C. Merriam Company, Springfield, Mass., p. 2121.

W. M. Gaines and P. P. Fischer: We wish to thank Mr. Walton for his active interest in the specifications of performance of analog-to-digital converters. The purpose of the paper was to stimulate greater interest in the standardization in the area of analog-to-digital converter specification, and if we can get positive constructive effort initiated our objectives have been met.

However, we take issue with Mr. Walton's statement that there are well-established definitions for specifying performance of analog-to-digital converters. One of the problems which faced us was utilizing established definitions within the instrumentation and electronic industries to meet the special needs of the analog-to-digital converter branch of instrumentation. We found wide divergence of wording and, in some cases, even intent. While we do not wish to engage in a quoting contest it is of interest that the definition of "accuracy in measurement" given in the "International Dictionary of Physics and Electronics"[1] is "The degree of correctness with which a method of measurement yields the 'true' value of a measured quantity. It is usually expressed in terms of error, the units being those of the measured quantity or the ratio (or per cent) of the error to the full scale value or to the actual value."

The wording of this definition differs in detail from both Pender and Del Mar[2] and the paper. Actually, both of the quoted definitions omit the extremely important reference to an absolute standard. The definition in the paper included the phrase "with reference to an accepted standard." With the exception of "drift," there seems to be no differences in intent in the other terms mentioned by Mr. Walton, merely a variation in wording.

A definition which has been used in the electronic amplifier field was adopted for drift and drift was related to the zero offset drift. Repeatability was considered to include the broader definition of drift quoted by Mr. Walton. Actually, the definition of drift given in the "International Dictionary of Physics and Electronics" is broader than in Considine.[3] In reference to "magnetic amplifiers," drift is defined as "...all effects other than changes in supply voltage, frequency, and ambient temperature." This definition therefore does not restrict the drift to constant ambient conditions.

All of this reduces to the simple fact that there is confusion in the definitions of the terms which constantly appear in the specification of performance of analog-to-digital converters. It is our hope that appropriate cognizant groups can initiate the necessary standards to improve this confusing situation.

REFERENCES

1. INTERNATIONAL DICTIONARY OF PHYSICS AND ELECTRONICS (book). D. Van Nostrand Company, Inc., Princeton, N. J., 1956.
2. See reference 1 of the discussion.
3. See reference 2 of the discussion.

Analysis of Conventional Single-Ended Magnetic Amplifiers with Inductive-Resistive Load Including the Derivation of Transfer Functions

NASSER E. NAHI
ASSOCIATE MEMBER AIEE

THE OPERATION of single-ended magnetic amplifiers with resistive load has been thoroughly explained in the literature and various types of analysis have been suggested.[1–5] However, no analytical method has yet been presented by which an input-output relation for single-ended magnetic amplifiers with inductive load can be obtained. In order to predict the value of the output corresponding to a given input, use of certain graphical techniques is introduced in reference 6, but this method fails to give any information about the transient response of the system.

In this paper the concept of a cycle of operation is introduced; it is defined as the interval bounded by two successive times when a core employed in the circuit initiates its resetting process. The cycle of operation does not necessarily coincide with, or even have a fixed position with respect to, a cycle of the supply voltage. The analysis method which is presented herein employs relations involving cycles of operations rather than half-cycles of the supply voltage. Therefore, it is possible to analyze circuits where commutation periods exist and the length of each period varies with the output quantity.

In this work single-ended magnetic amplifiers with resistive-inductive loads are analyzed. Operation of the doubler circuit with inductive load is explained and an analytical expression is developed by means of which the steady-state and transient response of the amplifiers can be predicted. The operation of a bridge circuit with inductive load and one with resistive load are quite similar because the load inductance is short-circuited by the rectifiers. Hence this circuit is only discussed briefly and the results are given. The center-tap circuit with inductive load is highly unstable for certain portions of the steady-state characteristic curve,[7] and will not be discussed in this paper.

Doubler Circuit with Inductive-Resistive Load

OPERATION OF THE AMPLIFIER

The amplifier circuit is shown in Fig. 1; R represents the total resistance of each gate circuit, i.e.,

$$R = R_L + R_g + R_r$$

where R_L is the load resistance, R_g the gate winding resistance, and R_r the equivalent forward resistance of each rectifier. The supply voltage v_g and the voltage across a separate winding with N_g turns on core *2* ($N_g\dot{\phi}_2$) are shown superposed in Fig. 2(A). In Fig. 2(B), i_{Lm} represents the load current if v_g were continuously applied to the load, i.e., if both cores would remain saturated. The value of the phase shift θ is given by the relation

$$\theta = \mathrm{tg}^{-1}\frac{Lw}{R}$$

With the polarities as shown in Fig. 1, core *1* will undergo a gating process during any positive half-cycle, e.g., the nth half-cycle, Fig. 2(A), and will eventually saturate during this period (the normal operation of the amplifier). Throughout this interval core *2* has been resetting at a relatively fast rate due to the coupling through the control windings.

When core *1* saturates ($t = t_2(n)$) the reset voltage applied to core *2* will drop con-

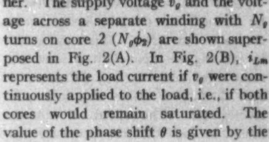


Fig. 1. Doubler circuit with inductive-resistive load

Paper 61-148, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted June 17, 1960; made available for printing December 1, 1960.

NASSER E. NAHI is with the University of Southern California, Los Angeles, Calif.

The work reported in this paper was performed in partial fulfillment of the requirements for the degree of doctor of philosophy at Purdue University, Lafayette, Ind., and was partially supported by the National Science Foundation. The author wishes to acknowledge the valuable guidance of his thesis supervisor, Dr. F. J. Friedlaender of Purdue University.

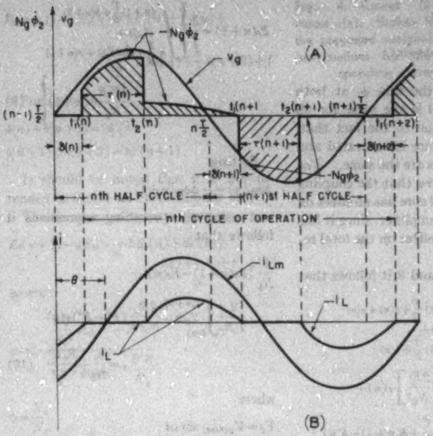


Fig. 2. (A) The supply voltage and $N_g\dot{\phi}_2$, and (B) maximum load current i_{Lm} and current i_L corresponding to $t_2(n)$

siderably since the voltage across the control winding of core *1* is now due only to the saturated inductance of its windings, which is very low. The load current begins to flow at $t=t_2(n)$ and will flow through the gate winding of core *1*. Since the current lags behind the voltage N_g, its flow continues through a portion of the negative half-cycle [the $(n+1)$st half-cycle] until it tends to change its direction. At this moment $[t=t_1(n+1)]$ rectifier *1* will block and rectifier *2* will conduct. The reset interval of core *2* ends at this time and its gating period starts immediately with the voltage $v_g-R_Li_2$ across the core (as will be shown later, i_2 is constant during this interval). Again this core saturates at $t_2\ (n+1)$ and remains saturated until the load current becomes zero. This process will repeatedly take place during both the transient and steady-state parts of the operation where, for the latter case,

$$t_1(n+1)=t_1(n)+\frac{T}{2}$$

$$t_2(n+1)=t_2(n)+\frac{T}{2}$$

Oscillograms of N_g, $N_g\dot{\phi}_2$, i_{Lm}, and i_L are shown in Fig. 3.

A cycle of operation is defined as the period bounded by two successive times when the core initiates its resetting interval. The cycle of operation pertaining to each core of the amplifier is divided into three intervals. The intervals related to core *2* and the corresponding sets of equations are described in the following. The reset portion of a general cycle of operation is denoted by (n) and the gating portion by $(n+1)$ since they are initiated during the nth and $(n+1)$st half-cycles respectively.

INTERVAL 1

This interval covers the period when both cores are unsaturated and core *2* is resetting, i.e., when $t_1(n)<t<t_2(n)$ (Fig. 2). During this time rectifier *1* is conducting and rectifier *2* acts like an open circuit.

The following relations expressing the core constraints and circuit equations are valid throughout this period.

$$N_ci_c+N_gi_1=F_g' \qquad \text{(core 1)}$$

$$N_ci_c+N_gi_2=-F_r' \qquad \text{(core 2)}$$

$$v_g=N_g\dot{\phi}_1+Ri_1+L\frac{di_1}{dt}$$

$$i_2=0$$

$$e_c=R_ci_c+N_c(\dot{\phi}_1+\dot{\phi}_2)$$

where F_g' and $-F_r'$ are the magnetomotive force (mmf) requirements of the gating and resetting core respectively. Further simplification is possible by assuming F_g' to be equal to F_r'. This is actually verified by experimental observations on the **B-H** loop and, furthermore, it can be justified by the fact that the voltages across the two cores during this interval are of the same magnitude as long as the control resistance is not very high. If the control circuit is highly constrained the mmf requirements of the cores cannot be predicted analytically by present knowledge of the magnetic materials. Therefore no attempt is made to derive a transfer function for this type of circuit.

The preceding set of equations are solved for $\dot{\phi}_1$ and $\dot{\phi}_2$. The results are

$$\dot{\phi}_1=\frac{v_g-R\dfrac{2F'}{N_g}}{N_g} \tag{1}$$

$$\dot{\phi}_2=\frac{e_c}{N_c}+R_c\frac{F'}{N_c^2}-\frac{1}{N_g}\left(v_g-R\frac{2F'}{N_g}\right) \tag{2}$$

INTERVAL 2

During this interval core *1* is saturated. This period ~~covers a portion of the nth~~ and also $(n+1)$st half-cycle of the supply voltage, i.e., when $t_2(n)<t<t_1(n+1)$ (Fig. 2). The coupling between the gate winding of core *1* (which is now saturated) and the control winding is represented by the voltage $N_c\dot{\phi}_{1s}$ in the control circuit.

The following equations are valid during this interval.

$$N_ci_c+N_gi_2=-F'' \qquad \text{(core 2)}$$

$$i_2=0$$

$$\dot{\phi}_1=\dot{\phi}_{1s}$$

$$e_c=R_ci_c+N_c(\dot{\phi}_2+\dot{\phi}_{1s})$$

Hence

$$\dot{\phi}_2=\frac{e_c}{N_c}+R_c\frac{F''}{N_c^2}-\dot{\phi}_{1s} \tag{3}$$

Normally, for very small values of R_c/N_c^2 the mmf requirements of the resetting core during this interval are not constant throughout the period. However, experimental observations show that the average value of this quantity over the interval is approximately constant and

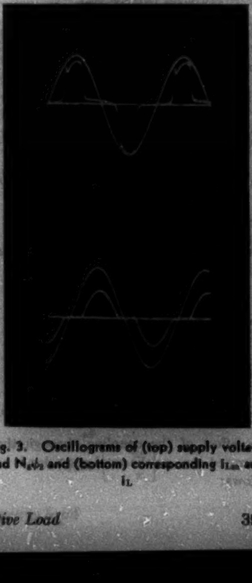


Fig. 3. Oscillograms of (top) supply voltage and $N_g\dot{\phi}_2$ and (bottom) corresponding i_{Lm} and i_L

equal to the corresponding value of the mmf of the core operating under higher control resistance conditions.

The stated experimental conclusion can be analytically verified for the limiting case $R_c/N_c^2 \approx e_c \approx 0$. With a first-order approximation, the mmf requirement of the core is given by the following expression:[8]

$$\text{mmf} = K_1 + K_2\dot{\phi}_2$$

For the case under consideration, equation 3 becomes

$$\dot{\phi}_2 = -\dot{\phi}_{1s}$$

As will be shown later, the average value of $\dot{\phi}_{1s}$ over the second interval is zero. Therefore, the average value of the mmf requirement becomes

$$\text{mmf}_{\text{avg}} = K_1$$

This implies that the average mmf is constant with a value which is independent of $\dot{\phi}_{1s}$.

INTERVAL 3

This interval covers a portion of the $(n+1)$st half-cycle namely, when $t_1(n+1) < t < t_2(n+1)$. Due to the symmetry of the circuit the instantaneous voltage across core *2* during this interval is the same as the voltage across core *1* during the first interval except for a change in the sign of v_g. Therefore it follows that

$$\dot{\phi}_2 = -\frac{1}{N_g}\left(v_g + R\frac{2F'}{N_g}\right) \qquad (4)$$

DERIVATION OF THE DETERMINING EQUATION OF THE SYSTEM

In order to simplify the manipulations involved, the following quantities are introduced:

$\delta(n) = t_1(n) - (n-1)\frac{T}{2}$ (Fig. 2)

$\Delta\phi_{2r}(n)$ = total reset flux change of core *2* during the nth interval

$\Delta\phi_{2g}(n+1)$ = total gating flux change of core *2* during the $(n+1)$st interval

$\tau(n) = t_2(n) - t_1(n)$ (Fig. 2)

The total reset flux change is given by the following relation:

$$\Delta\phi_{2r}(n) = \int_{(n-1)\frac{T}{2}+\delta(n)}^{(n-1)\frac{T}{2}+\delta(n)+\tau(n)} \dot{\phi}_2 dt + \int_{(n-1)\frac{T}{2}+\delta(n)+\tau(n)}^{n\frac{T}{2}+\delta(n+1)} \dot{\phi}_2 dt$$

The effect of the coupling through the saturated core on the total reset flux is as follows:

$$\Delta\phi_{1s} = \int_{(n+1)\frac{T}{2}+\delta(n)+\tau(n)}^{n\frac{T}{2}+\delta(n+1)} \dot{\phi}_{1s} dt \approx 0$$

since the value of the flux ϕ_1 at both limits of the integral is the same for all practical purposes due to the fact that, for both cases, the core is saturated and the mmf requirements are the same. The above conclusion shows that the coupling through the saturated core has no effect on the operation of the amplifier, since it has relatively very small effect on the total reset flux change.

From equations 2 and 3 it follows that

$$\Delta\phi_{2r}(n) = -\frac{1}{N_g}\int_{(n-1)\frac{T}{2}+\delta(n)}^{(n-1)\frac{T}{2}+\delta(n)+\tau(n)} V_g dt + \left[R_c\frac{F-F''}{N_c^2} + R\frac{2F'}{N_g^2}\right]\tau(n) + \left[\frac{e_c}{N_c} + R_c\frac{F''}{N_c^2}\right]\left[\frac{T}{2} + \delta(n+1) - \delta(n)\right] \qquad (5)$$

If the transient response of the amplifier is not very fast, $\delta(n+1) - \delta(n)$ is negligible as compared with $T/2$. Therefore,

$$\Delta\phi_{2r}(n) = -\frac{1}{N_g}\int_{(n-1)\frac{T}{2}+\delta(n)}^{(n-1)\frac{T}{2}+\delta(n)+\tau(n)} V_g dt + \frac{e_c}{N_c}\frac{T}{2} + \left[\frac{R_c}{N_c^2}(F'-F'') + 2R\frac{F'}{N_g^2}\right]\tau(n) + \frac{R_cF''}{N_c^2}\frac{T}{2} \qquad (6)$$

$\Delta\phi_{2g}(n+1)$ is obtained from equation 4.

$$\Delta\phi_{2g}(n+1) = -\frac{1}{N_g}\times \int_{n\frac{T}{2}+\delta(n+1)}^{n\frac{T}{2}+\delta(n+1)+\tau(n+1)} V_g dt - 2R\frac{F'}{N_g^2}\tau(n+1) \qquad (7)$$

It should be noted that for very small values of R_c/N_c^2, F'' represents the average value of the mmf requirement of the resetting core during the second interval. The magnitude of the average output voltage during the nth and $(n+1)$st intervals are respectively

$$E_0(n) = \frac{2}{T}\left\{\int_{(n-1)\frac{T}{2}+\delta(n)+\tau(n)}^{n\frac{T}{2}+\delta(n+1)} V_g dt + \frac{2F'R}{N_g}\tau(n)\right\} \qquad (8)$$

$$E_0(n+1) = \frac{2}{T}\left\{\int_{n\frac{T}{2}+\delta(n+1)+\tau(n+1)}^{(n+1)\frac{T}{2}+\delta(n+2)} V_g dt + \frac{2F'R}{N_g}\tau(n+1)\right\} \qquad (9)$$

Setting

$$\Delta\phi_{2r}(n) = -\Delta\phi_{2g}(n+1)$$

and using the preceding expressions it follows that

$$\frac{1}{N_g}[E_0(n+1) - E_0(n)] + \frac{2}{TN_g}\int_{\delta(n)}^{\delta(n+2)} V_g dt - \frac{2R_c}{N_c^2T}(F'-F'')\tau(n) = \frac{e_c}{N_c} + \frac{R_cF''}{N_c^2} \qquad (10)$$

where

$$V_g = V_{g(\max)}\sin\omega t$$

In general, if the control voltage is given by

$$e_c = E_{c\max}\sin\Omega t,\ (\Omega \ll \omega)$$

the quantity e_c in equation 10 should be replaced by its average value:

$$E_c(n) = \text{avg } e_c = \frac{2}{T}\int_{(n-1)\frac{T}{2}+\delta(n)}^{n\frac{T}{2}+\delta(n+1)} e_c dt$$

In order to derive a linear expression for $\tau(n)$ in terms of $E_0(n)$ the cosine term in equations 8 and 10 can be approximated by a straight line such as given in equation 11.

$$\cos\omega t \approx 1.25\left(1 - \frac{2\omega}{\pi}t\right) \qquad 0 < t < T/2 \qquad (11)$$

This straight-line approximation will introduce a maximum error of less than 9% and an rms error of less than 5% in the actual value of $\cos\omega t$ for a portion of input-output characteristic which corresponds to up to approximately 96% of the full output of the system.

It should be noted that the cosine term can be approximated by other straight lines as well. In fact, for relatively small variations of input the straight-line tangent to the cosine curve at the point corresponding to the operating point on the characteristic curve yields better results. The approximation would have been totally unnecessary if a supply voltage with a square waveform was employed instead of the sinusoidal voltage, since equations 8 and 9 would have the transcendental form.

Eliminating $\tau(n)$ between equations 8 and 10 results in the following expression:

$$E_0(n+1)-\alpha E_0(n)=bE_c(n)+\frac{d}{2}\delta(n)+\frac{d+h}{2}\delta(n+1)+\frac{h}{2}\delta(n+2)+f$$

Furthermore,

$$\delta(n)+\delta(n+1)=2\delta'(n)$$
$$\delta(n+1)+\delta(n+2)=2\delta'(n+1)$$

It should be noted that it is not intended to derive an expression for $\delta(n)$.

The above substitutions result in

$$E_0(n+1)-aE_0(n)=bE_c(n)+d\delta'(n)+h\delta'(n+1)+f \quad (12)$$

where

$$a=1-\frac{\frac{N_g}{N_c^2}R_c(F'-F'')}{\frac{2.5}{\pi}V_{g(\max)}-2\frac{F'R}{N_g}}$$

$$b=\frac{N_g}{N_c}$$

$$d=\frac{5\omega V_{g(\max)}}{\pi^2}+\frac{2}{T}\frac{N_gR_c}{N_c^2}(F'-F'')\times\frac{\frac{5}{\pi}V_{g(\max)}}{-\frac{2.5}{\pi}V_{g(\max)}+\frac{2F'R}{N_g}}$$

Since

$$\frac{F'R}{N_g}\ll V_{g(\max)}$$

then

$$d\approx\frac{5\omega V_{g(\max)}}{\pi^2}-\frac{4}{T}\frac{N_gR_c}{N_c^2}(F'-F'')$$

$$h=-\frac{5\omega V_{g(\max)}}{\pi^2}$$

$$f=\frac{R_cF''}{N_c^2}N_g-\frac{R_cN_g}{\pi N_c^2}(F'-F'')\times\frac{2.5V_{g(\max)}}{-\frac{2.5}{\pi}V_{g(\max)}+\frac{2F'R}{N_g}}$$

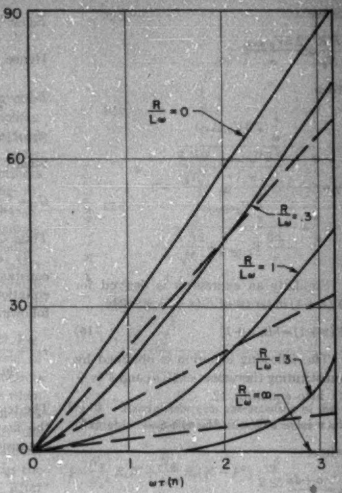
In the preceding difference equation $\delta'(n)$ and $\delta'(n+1)$ are functions of $E_0(n)$ and $E_0(n+1)$ respectively. In the following, expressions for $\delta'(n)$ and $\delta'(n+1)$ are derived.

$t_1(n+1)$ is the time during the $(n+1)$st half-cycle when the load current becomes zero, i.e., the solution of the following equation:

$$\sin(\omega t-\theta)-\left\{\sin[\omega(t(n)+\delta(n))-\theta]-\frac{2F'}{N_gI_{g(\max)}}\right\}\exp\left[-\frac{R}{L}(t-\tau(n)-\delta(n)\right]=0 \quad (13)$$

From Fig. 2 it is seen that the solution of this equation is

Fig. 4. Curves $(\theta-\delta'(n))$ versus $\tau(n)$. Broken lines are the suggested straight-line approximations for the corresponding curves



$$t=T/2+\delta(n+1)$$

$\delta(n)$ and $\delta(n+1)$ are replaced by their average value $\delta'(n)$. This is justified by the fact that $\delta(n)$ varies slowly with respect to n. Under this condition equation 13 can be solved for $\delta'(n)$. The result is

$$\delta'(n)=\frac{\theta}{\omega}-\frac{1}{\omega}t_g^{-1}\times\frac{\sin\omega\tau(n)}{e^{R/L\left[\frac{T}{2}-\tau(n)\right]}+\cos\omega\tau(n)}+\frac{1}{\omega}\sin^{-1}\left\{\frac{2F'}{N_gI_{(\max)}}\times\sqrt{\frac{1}{\sin^2\omega\tau(n)+\left[e^{R/L\left[\frac{T}{2}-\tau(n)\right]}+\cos\omega\tau(n)\right]^2}}\right\}$$

The magnitude of the third term in the above equation is usually very small (since $F'/N_g\ll I_{\max}$ and the quantity under the radical is normally less than 1) and the term can be ignored. Since the simplified equation is still nonlinear, if it is to be used in its present form prediction of the response of the amplifier would require application of some numerical methods. However, the above equation can be approximated by the following straight line:

$$\delta'(n)=\frac{\theta}{\omega}-K\tau(n) \quad (14)$$

where K, the slope of the straight line, is obtained by making the average error of the quantity $\delta'(n)$ zero over the interval of interest. The curves $\delta'(n)$ versus $\tau(n)$ are only functions of the parameter R/L, hence the quantity K can easily be obtained by plotting the curve for the desired value of R/L and choosing K so that the areas under the curve and the straight line are equal. However, this can also be done analytically by finding the magnitude of the function at m arbitrary points by substituting the m values of $\tau(n)$ in the equation of the curve. In order to show the magnitude of the error involved due to this approximation, the curves $\delta'(n)$ versus $\tau(n)$ for different values of R/L are shown in Fig. 4. This error approaches zero for both low and high values of R/L. In the following K is chosen to equal θ/π, which results in zero error for the boundary values of R/L, namely, zero and infinity.

Substituting the value $k=\theta/\pi$ into equation 14 and applying the results to equations 8 and 11 results in the following:

$$\delta'(n)=\lambda E_0(n)+\gamma \quad (15)$$

where

$$\lambda=\frac{-\frac{T}{2}}{\left(\frac{5}{\pi}-\frac{2.5}{\theta}\right)V_{g(\max)}-\frac{2\pi}{\theta}\frac{F'R}{N_g}}$$

or

$$\lambda\approx\frac{\pi\theta}{2.5\omega V_{g(\max)}\left[1-\frac{2\theta}{\pi}\right]}$$

Since

$$\frac{F'R}{N_g} \ll \frac{2.5V_{g(\max)}}{\pi}$$

then

$$\gamma = \frac{\dfrac{\pi}{\omega} + \dfrac{\dfrac{1}{\omega}(2.5V_{g(\max)})}{-\dfrac{2.5V_{g(\max)}}{\pi} + \dfrac{2F'R}{N_g}}}{\dfrac{\pi}{\theta} + \dfrac{\dfrac{5}{\pi}V_{g(\max)}}{-\dfrac{2.5}{\pi}V_{g(\max)} + \dfrac{2F'R}{N_g}}} \approx 0$$

Similarly an expression is derived for $\delta'(n+1)$ in terms of $E_0(n+1)$, namely

$$\delta'(n+1) = \lambda E_0(n+1) \tag{16}$$

The following equation is obtained by substituting the values of $\delta'(n)$ and $\delta'(n+1)$ into equation 12.

$$E_0(n+1) - \alpha E_0(n) = \beta E_c(n) + \xi \tag{17}$$

where

$$\alpha = \frac{a + d\lambda}{1 - h\lambda}$$

$$\beta = \frac{b}{1 - h\lambda}$$

$$\xi = \frac{f}{1 - h\lambda}$$

For low values of θ

$$h\lambda = \frac{-2\theta}{\pi\left(1 - \dfrac{2\theta}{\pi}\right)} \ll 1$$

then

$$\beta = \frac{N_g}{N_c}$$

$$\xi = -\frac{5V_{g(\max)}}{N_g}$$

It should be noted that since i_L is approximately zero at $t_2(n)$ and $t_1(n+1)$ (Fig. 2) then

$$E_0(n) = RI_L(n)$$

where

$$I_L(n) = \frac{2}{T}\int_{(n-1)\frac{T}{2}+t_1(n)}^{n\frac{T}{2}+t_1(n+1)} i_L dt$$

DERIVATION OF THE TRANSFER FUNCTION

Solution of equation 17 can be obtained (for both d-c and a-c control signals) by directly solving the difference equation. However, in the following the solution will be obtained by operational calculus.[10]

The following expression is obtained by taking the Laplace transformation of equation 17.

$$E_0(s)e^{(\pi/\omega)s} - \alpha E_0(s) = \beta E_c(s) + \frac{\xi}{s}$$

Hence

$$E_0(s) = \frac{1}{e^{(\pi/\omega)s} - \alpha}\left[\beta E_c(s) + \frac{\xi}{s}\right] \tag{18}$$

STEP FUNCTION RESPONSE

Steady-State Incremental Gain:

$$G_{ss} = \frac{\beta}{1 - \alpha} \tag{19}$$

Time Constant:

By applying the definition of time constant for the first-order systems to equation 18 with $E_c(n) = \text{constant}$, the following results:

$$\tau = -\frac{1}{\ln \dfrac{1}{\alpha}} \tag{20}$$

The term $\ln 1/\alpha$ can be approximated by the first two terms of its Taylor series expansions ($\alpha \approx 1$). The result is

$$\tau = \frac{\pi}{\omega(1 - \alpha)}$$

SINUSOIDAL RESPONSE

If the control voltage is given by

$$e_c = E_{c(\max)} \sin \Omega t$$

then

$$E_c(n) = \text{half-cyclic avg } e_c$$

$$= \frac{2}{T}\int_{(n-1)\frac{T}{2}+\delta(n)}^{n\frac{T}{2}+\delta(n+1)} e_c dt$$

Furthermore, since $\Omega \ll \omega$ and

$$\delta(n+1) - \delta(n) \ll \frac{T}{2}$$

then

$$E_c(n) \approx \frac{2}{T}\int_{(n-1)\frac{T}{2}}^{n\frac{T}{2}} e_c dt$$

The above equation can be written as follows:

$$E_c(n) = \frac{2}{T}\left\{\int_0^{n\frac{T}{2}} e_c dt - \int_0^{(n-1)\frac{T}{2}} e_c dt\right\}$$

The Laplace transformation of the above equation is

$$E_c(s) = \frac{\omega}{\pi}[-e^{-\pi/\omega s} + 1]\mathcal{L}\int_0^{(n-1)\frac{T}{2}} e_c dt$$

Identifying $nT/2$ (the discrete variable) with the operational[4] variable t and then taking the Laplace transformation of

$$\int_0^t e_c dt$$

results in the following:

$$E_c(s) = \frac{\omega}{\pi}\left\{\frac{1}{s}[1 - e^{-\pi/\omega s}]\right\}E_{c0}(s)$$

where

$$E_{c0}(s) = \mathcal{L}[e_c]$$

Hence the following relation (in frequency domain) between the input and output is derived from equation 18.

$$E_0(s) = \frac{1}{e^{\frac{\pi}{\omega}s} - \alpha}\frac{2\beta}{T}\frac{1}{s}(1 - e^{-\frac{\pi}{\omega}s})E_{c0}(s) + \frac{\xi}{s} \tag{21}$$

The inverse transform of $E_0(s)$ (equation 21) is in general a discontinuous function of time. This solution is an interpolation (a sinusoidal interpolation in case of the steady-state operation) of the the discrete solution of the difference equation 17. Furthermore, this solution is the same as that of the modular differential equation corresponding to the difference equation under consideration, as introduced and discussed in reference 5.

BLOCK DIAGRAM REPRESENTATION

A block diagram representing the amplifier can be derived from equations 11, 15, and 16. This diagram (Fig. 5) plainly shows the effect of the inductive load on the system by the negative feedback loop.

It should be noted that only the parameter λ which appears as the coefficient of the feedback element is a function of θ (equation 15). For the marginal value of $L/R = 0$ (i.e., the case of the resistive load) λ, and therefore the amount of feedback, is zero. The simplified result, which is the block diagram representation of the amplifier with resistive load, agrees

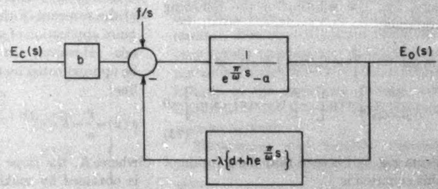


Fig. 5. Block diagram representation of the amplifier

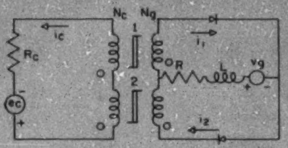


Fig. 6. Bridge circuit with inductive-resistive load

with the results obtained in the literature,[2] as expected.

In the following the incremental magnitude ratio and the phase lag refer to the interpolate curve defined above.

Incremental Magnitude Ratio:

$$\frac{\beta}{\lambda}\left|\frac{1-e^{-j\lambda}}{-\alpha+e^{j\lambda}}\right|=\frac{\beta}{\lambda}\sqrt{\frac{2(1-\cos\lambda)}{1+\alpha^2-2\alpha\cos\lambda}} \qquad (22)$$

where

$$\lambda=\pi\frac{\Omega}{\omega}$$

Phase Lag:

$$\frac{\pi}{2}+\left\{\underline{/-\alpha+e^{j\lambda}}-\underline{/1-e^{-j\lambda}}\right\}$$

$$=\frac{\pi}{2}+\lambda-\tan^{-1}\frac{(1-\alpha)\sin\lambda}{(1+\alpha)(\cos\lambda-1)} \qquad (23)$$

The time constant of the system is obviously the same as that obtained for a step of input in the previous section.

Fig. 7. Oscillograms of (top) $N_g\phi_g$ and v_g (bridge circuit) and (bottom) i_L

Bridge Circuit with Inductive-Resistive Load

PRINCIPLE OF OPERATION

The principles of operation of a conventional bridge magnetic amplifier with resistive and with inductive load are the same because in the latter case the load is short-circuited by rectifiers and no commutation can occur (Fig. 6). Therefore, the cycle of operation of the system coincides with a cycle of the supply voltage. The load current i_L flows continuously and, normally, has a value larger thanthe excitation current flowing through the gate circuit of the gating core (when it is not saturated).

With polarities as illustrated in Fig. 6, and during a positive half-cycle of the supply voltage, rectifier *2* is conducting throughout a period which ends when core *1* saturates (obviously rectifiers *1* and *3* are conducting throughout the positive half-cycle). Therefore, almost all of the voltage v_g is applied across the winding of core *1* until it saturates. After this transition, current i_1 will be large enough to cause rectifier *2* to block, which will remain so until the end of the half-cycle. Fig. 7 illustrates oscillograms of $N_g\phi_2$ and v_g superposed, and the load current i_L.

DETERMINING EQUATION OF THE SYSTEM

The general method of analysis is applied directly to this system. Since the operation of this amplifier with inductive-resistive load very closely resembles its operation with resistive load (which has been discussed in the literature[3,4]) only the results of the analysis are given below.

The half-cyclic average values of input and output satisfy the following relation:

$$E_0(n+1)-\alpha E_0(n)=\beta E_c(n)+\xi$$

where

$$E_0(n)\doteq\frac{2}{T}\int_{(n-1)\frac{T}{2}}^{n\frac{T}{2}}e_c dt$$

$$E_c(n)\doteq\frac{2}{T}\int_{(n-1)\frac{T}{2}}^{n\frac{T}{2}}e_c dt$$

and

$$\alpha=1-\frac{\frac{N_g}{N_c^2}R_c(F'-F'')}{\frac{2.5}{\pi}V_{g(\max)}}$$

$$\beta=\frac{N_g}{N_c}$$

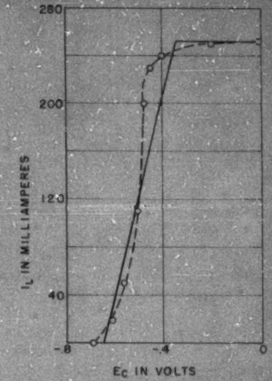


Fig. 8. Steady-state gain characteristic of the doubler amplifier with $R_c=533$ ohms

$$\xi=\frac{N_g}{N_c^2}R_cF''+\frac{N_g}{N_c^2}R_c(F'-F'')\frac{2.5}{2.25}$$

Again a transfer function for the amplifier can be derived from the preceding difference equation, which will have the same form as that of a doubler circuit except for the different values of the parameters given above.

Experiments

The circuit arrangement of Fig. 1 was employed with the following components. Cores were 50% nickel-iron (Orthonol) supplied by Magnetics Incorporated, catalog number *50020* (tape thickness, 2 mils; inside diameter, 1.625 inches; outside diameter, 2 inches; height, 1 inch; mean length of magnetic path, 14.49 centimeters; effective core cross-sectional area, 1.030 centimeters[2]). Rectifiers were

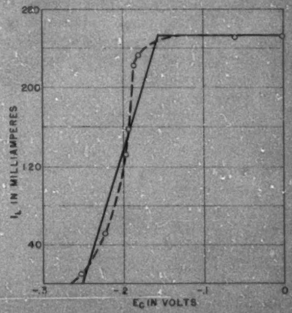


Fig. 9. Steady-state gain characteristic of the doubler circuit with $R_c=133$ ohms

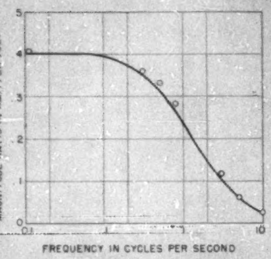


Fig. 10. Frequency response of the doubler amplifier with $R_c = 533$ ohms

General Electric Company type *538* silicon junction rectifiers, 250-milliampere rating.

The pertinent parameters of the circuit were

$N_g = 1,500$ turns
$N_c = 500$ turns
$L = 0.47$ henry
$v_g = 55$ volts rms
$R =$ as shown on the figures

Figs. 8 and 9 illustrate the steady-state gain characteristic of the amplifier for different values of R_c. The experimental results are shown by the dotted lines and the analytical results by solid lines.

Fig. 10 illustrates the frequency response of the amplifiers.

Conclusions

This paper has demonstrated the feasibility of deriving transfer functions to represent the static and dynamic behavior of magnetic amplifiers with resistive-inductive loads. The operation of single-ended magnetic amplifiers with inductive loads has been analyzed and corresponding transfer functions have been developed. It is shown that the analytical and experimental results are in close agreement.

References

1. Analysis of Magnetic Amplifiers by Use of Difference Equations, P. P. Johannessen. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 73, 1954 (Jan. 1955 section), pp. 700–11.

2. Dynamic Operation of Magnetic Amplifiers with Real-Core and Rectifier Functions, H. C. Bourne, Jr., D. Nitzan. *Ibid.*, vol. 76, Nov. 1957, pp. 646–55.

3. Magnetic Amplifier Analysis and Application Using Block-Diagram Techniques, Robert M. Hubbard. *Ibid.*, pp. 578–88.

4. Analysis of the Full Wave Magnetic Amplifier Circuits Considering the Change of the Width of the Dynamic Hysteresis Loop, Tadishi Kikuchi. *Ibid.*, July, pp. 241–49.

5. The Analysis of Magnetic Amplifiers, Part I—Single Feedbacks, L. A. Finzi, J. J. Suozzi. *Ibid.*, vol. 77, 1958 (Jan. 1959 section), pp. 1019–31.

6. The Analysis of Magnetic Amplifiers with Various Types of Load, Part I—Load Currents for Given Angle of Firing, L. A. Finzi, R. R. Jackson. *Ibid.*, vol. 73, July 1954, pp. 270–78.

7. A General Method for Analyzing Magnetic Amplifiers Including the Derivation of Transfer Functions, N. E. Nahi. *Ph.D. Thesis*, Purdue University, Lafayette, Ind., 1960.

8. Core-Reset Functions in Magnetic Amplifier Analysis, Part I—A Core Reset Function, G. C. Feth. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Sept. 1958, pp. 503–09.

9. Flux Reversal in Magnetic Amplifier Cores, F. J. Friedlaender. *Ibid.*, vol. 75, July 1956, pp. 268–78.

10. Operational Calculus Based on the Two-Sided Laplace Transform (book), V. D. Pol. Cambridge University Press, Cambridge, England, 1955.

Problems Encountered in Applying the Silicon-Controlled Rectifier to Control Tungsten Lamp Loads

C. S. DAUGHERTY
MEMBER AIEE

THIS PAPER will show the problems encountered in using silicon-controlled rectifiers (SCR) as control devices to vary the light output of tungsten lamp loads such as are commonly used in theater or studio stage lighting. A particular application using magnetic amplifier control of the SCR will be discussed as a lamp dimmer.

Historically, tungstem lamp dimmers have been built using (1) variable series resistance, (2) series saturable reactors, (3) series magnetic amplifiers, (4) variable-voltage autotransformers, (5) thyratron tubes, and now (6) the silicon-controlled rectifier. The relative merits of these devices will be discussed later.

Applying the SCR

The SCR[1] is a semiconductor device which functions like a thyratron. Once triggered "on" into a forward conducting state, it will remain "on" until the conducting current falls below a threshold value, at which time it becomes nonconducting and will block the flow of forward current. The time in the applied voltage cycle at which the device is triggered can be controlled, and, thus, variable output voltage can be applied to the lamp load.

The SCR, as any silicon semiconductor, has its thermal and voltage potential limitations, and therefore is sensitive to forward surge currents and transient inverse voltage peaks.[2] For continuous operation at a rating within its design it must be properly heat-sinked to allow proper stud cooling.[3] The tremendous surge currents resulting from voltage suddenly applied to a cold tungsten filament are particularly troublesome to this device. Peak currents as high as 20 times the steady-state rating may result. Present production SCR's have a nonrecurrent 1/2 cycle (60-cycle basis) surge rating of approximately three times the steady-state rating.

Fig. 1 shows a workable tungsten lamp dimmer circuit employing SCR's. This is a full-wave a-c output circuit. The power-controlling devices are the SCR's, shown here as SCR*1* and SCR*2*, arranged so that they fire on alternate half-cycles. The firing angle of the SCR's is controlled by a magnetic amplifier, consisting of a split-output common-control circuit. The output windings are fed from a circuit which effectively isolates the gate cathode. This prevents extraneous firing of the SCR's due to interaction. A magnetic amplifier is used for gate firing because with it the SCR's can be fired from many piled-on signals, a common practice in theater lighting control.

The control for this device is taken from a half-wave rectified source. This is commonly used in the lighting control

---

Paper 61-182, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 16, 1959; made available for printing December 6, 1960.

C. S. Daugherty is with Vickers Inc., St. Louis, Mo.

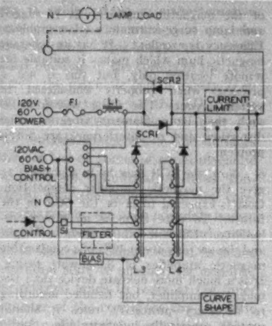


Fig. 1. Typical circuit

industry in presetting systems using half-wave rectifiers as blocking and switching elements. The presetting system establishes preset light intensities in programs. Using half-wave control for a full-wave magnetic amplifier requires a filter network to make the unit insensitive to the phase of control source. This is necessary since the control power may be taken from any phase of the 3-phase supply and not necessarily from the same phase as the SCR power. Fig. 2 shows the result of this filter on various supply voltage phases. The curves show the maximum difference in output voltage for the same control signal for any phase condition. The minimum output voltage is controlled by biasing the control reactors to a point that will not fire the SCR's. An extremely linear curve of output voltage versus control position is obtained using negative feedback. Fig. 3 shows the effect of various feedback quantities.

The circuit described is all that would be necessary if the SCR's did not require surge current suppression. In order to be able to turn on a lamp such as a 3,000-watt senior spot without exceeding the permissible transient current ratings of the 16-ampere SCR, a current-limit circuit is necessary. The action of this circuit is such that as the current rises above a predetermined level, a reference voltage that has been blocking this circuit lets a higher signal voltage pass, desaturating the magnetic amplifier cores and causing the SCR's to return to a nonconducting state, thus limiting load current. This circuit will protect for surges longer than 1/2-cycle duration, and is particularly applicable to lamps such as the high-intensity spots where the warming time of the lamp requires approximately 60 cycles to reach steady-state current.

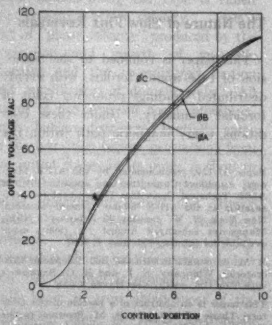


Fig. 2. Output voltage versus control position for various supply voltage phases

Since the time when the dimmer is to be turned on in the applied-voltage cycle cannot be predicted, and the SCR's can be in the "on" condition at the instant of turn-on due to the preset condition of the control magnetic amplifier cores, and since surge currents larger than the rating of the SCR can occur, the following method must be used to turn on the dimmer. The power to the control magnetic amplifier and its bias are turned on first, thus leaving the control magnetic amplifiers setting at cutoff and the SCR's in the nonconducting state. Power to the SCR's and control circuit is then turned on and voltage is applied to the lamps at a rate equal to the control-magnetic-amplifier time constant. This gives the above current-limit circuit the 1/2-cycle time necessary to function properly.

To protect the SCR's against short circuits and other overcurrent faults of less than 1/2-cycle duration, a fuse is used. This is a special fast-acting fuse and has I^2t rating less than that of the SCR's. It is a measure of the instantaneous heating of the SCR's and the fuse. Thus, the fuse will open before the instantaneous maximum heating rating of the SCR's is reached. (This fuse is now manufactured as a standard type.) Further considerations when applying SCR's to lamp loads must be given to audio noise generation within the SCR's, the noise generated by lamp filament vibration due to the fast-firing (microsecond range) of the SCR's, and the effect of voltage pulses on the supply line which are caused by the sudden firing of the SCR's.

By inserting a small series choke in the supply line the firing time is sufficiently increased in order that the filament vibration is within acceptable noise limits, and the voltage pulses on the supply line are reduced from 60% to less than 3% of the supply peak voltage, a harmless value on any known installation. Physically this choke is the size of a 20-va (volt-amperes) 60-cps (cycle-per-second) transformer on the 3.0-kw rating, and a 70-va transformer on the 10-kw rating. Noise measurements in an anechoic chamber show the sound pressure level of the dimmer to be highest in the 600–4,800-cps frequency range. In this frequency range sound pressure levels of less than 45 db (decibels) were measured. Also, the sharp firing causes radiated r-f (radio-frequency) interference. This has been eliminated by the use of a 0.05-mfd (microfarad) capacitor across the input line.[4]

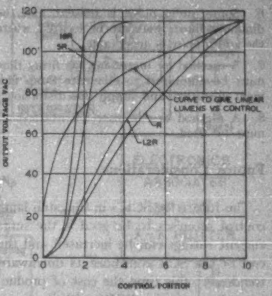


Fig. 3. Output voltage versus control position with various feedback quantities

In lighting work it is necessary to be able to connect a cold lamp load into a live control circuit. This is known as "hot patching." Hot patching on this unit is accomplished by a sliding interlock. It opens the control circuit and reduces the output voltage to near zero volts while load is added. The efficiency of this unit is excellent; it is 98%. Power factor of the unit with full-load rating is 96%.

Conclusions

The SCR's can be successfully applied to the control of tungsten lamp loads with the following precautions:

1. A proper current-limit circuit to reduce cold lamp surge current to the rating of the SCR must be used.
2. A time delay interlock must be used in turning on the dimmer installation.
3. To protect against short circuit of the output voltage a one-time fast-acting fuse must be used.
4. A special interlock must be used for hot patching.
5. Proper heat sink rating must be used for the wattage desired.

6. Some audio noise is present and the dimmers must therefore be located where the audio noise is acceptable.

7. A method for increasing the firing time must be employed to eliminate lamp filament vibration, and supply line distortion.

8. A method to eliminate r-f interference must be used.

Future Considerations

The future for SCR's in tungsten lamp control promises to be good if the surge current ratings can be increased and the cost of the SCR continues its downward trend. At this time the cost of producing a dimmer using SCR's with its necessary protective circuits is comparable to a magnetic amplifier power circuit at a 3,000-watt rating, and about 40% higher in a 10,000-watt size. In time the SCR price is expected to go down markedly while the magnetic amplifier will remain relatively constant. Thus SCR unit Production will be cheaper. The present cost is based on a 16-ampere SCR costing $20, and a 50 ampere SCR costing $75.

Comparison with Other Dimming Devices

Various devices have been used to control tungsten lamp voltage:

1. Variable series resistance (rheostat). This device is used very little today because of its low efficiency, poor load regulation, and because it requires maintenance.

2. Variable-voltage autotransformers. This device is used extensively in small hand-operated jobs and for a few larger hand-operated systems. Efficiency is excellent, regulation is good, and cost is low. It may be used for remote-controlled jobs with motor positioners, but the motorized unit costs nearly as much as magnetic amplifier or SCR units. Response of the motorized unit is generally slow and the unit requires continued maintenance. The noise level is lower than in the SCR unit or the magnetic amplifier unit, except when the motor is running to reposition the unit.

3. The saturable reactor is a slow and heavy, though highly reliable, unit. It has been generally replaced by the thyratron tube or magnetic amplifier system. The unit has the same magnetic hum as the magnetic amplifier system.

4. The thyratron tube system is an excellent remote control system, but the short lifespan of the power tubes is a costly replacement item. Efficiency is low compared with the magnetic amplifier or SCR. The problems associated with the sharp firing angle have to be considered in the same way as described for the SCR.

5. The magnetic amplifier system, although slightly heavier and more costly than the thyratron tube system, has proven to be the most reliable, long-life system on the market today. No other remote control system today has the proved reliability of the magnetic amplifier. It is rugged and lamp surge currents are no problem. Efficiency is excellent. It has an inherent magnetic hum which makes it suitable for remote location only, but this is not a problem when properly considered in system layout. The firing wavefront is such that lamp filament vibration, line distortion, and r-f interference are not a problem.

6. Now, the SCR may be in a position to give the magnetic amplifier competition. Cost is expected to favor the SCR in the future. To date there is no proved reliability. Efficiency is excellent. Weight and size are only about 16% of a comparable magnetic amplifier of 3,000-watt rating. It is a much more delicate device than the magnetic amplifier, but handled according to necessary protective rules it should provide practically indefinite life. Audio noise is about 20 db lower than in magnetic amplifier dimmers; it occurs in a higher frequency band, thus facilitating acoustic isolation.

References

1. Controlled Rectifier Manual (book). General Electric Company, Liverpool, N. Y., first edition, 1960.

2. Rectifier Voltage Transients: Their Generation, Detection and Reduction, F. W. Gutzwiler. General Electric Company.

3. Heatsinks for Stud-Mounted Semiconductors, F. W. Gutzwiler. General Electric Company, Auburn, N. Y.

4. G. E. Circuit Designers Guide (book). General Electric Company, Syracuse, N. Y., *ECG-425*.

Detailed Measurements of Slow Magnetization Processes in Tape-Wound Cores

R. M. BROWNELL
NONMEMBER AIEE

R. C. BARKER
ASSOCIATE MEMBER AIEE

THE TECHNOLOGY based on the saturable-core reactor, including magnetic amplifiers and other nonlinear magnetic devices,[1] has grown rapidly in the past decade. This has stimulated development of commercially available polycrystalline magnetic materials, and prompted several studies of their properties going well beyond the mere collection of empirical, macroscopic data. A recent paper,[2] an attempt to summarize the state of the art from the electrical engineering standpoint, indicated how changes in the magnetic state of a tape-wound core may take place. A lengthy bibliography included papers on the structure and arrangement of domain walls in tapes, calculations of their growth and propagation energies, and several papers showing microscopic domain patterns in isolated samples.

The group doing research on magnetic materials at Yale University has been performing the complementary task of setting down those macroscopic properties that need explanation. The effort during the past 2 years has been directed mainly toward obtaining more exact measurements of how commercial-grade tape cores behave under various conditions of excitation. Through these exploratory measurements it may be possible to give a better account of the relations between domain nucleation, growth, and annihilation and the macroscopic magnetic properties of these cores, and of the influence of their metallurgical and geometrical properties. It is the purpose of this paper to present the results of some of these measurements, taken at low rates of change of flux.

The Nature of Slow Flux Reversals

This paper is confined to the discussion of tape-wound toroids, with evenly distributed windings producing fields of toroidal symmetry. Under these conditions, every magnetic path within the

---

Paper 61-129, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 12, 1960; made available for printing February 8, 1961.

R. M. Brownell is with the Bell Telephone Laboratories, Whippany, N. J., and R. C. Barker is with Yale University, New Haven, Conn.

This paper is an abstract of a portion of the Doctoral Thesis submitted by R. M. Brownell to the Yale School of Engineering in partial fulfillment of the requirements for the Doctor of Engineering Degree. The research was partly supported by Magnetics, Inc., Butler, Pa.

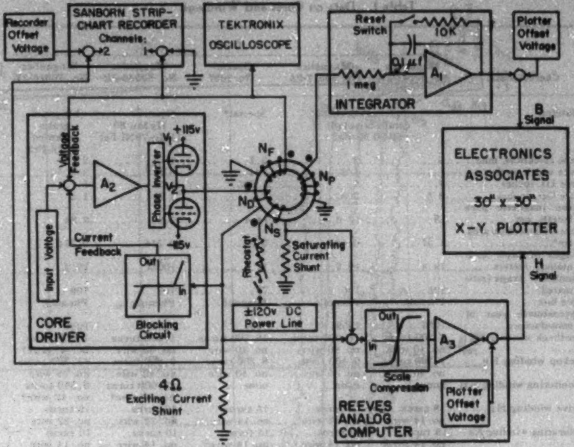


Fig. 1. Arrangement of equipment for measurements

A_1—KinTel model 111BFO low-drift d-c amplifier with no internal feedback
A_2—D-C amplifier, voltage gain approximately 1,000
A_3—Low-drift d-c amplifiers in computer
N_D—Drive winding
N_F—Feedback winding
N_P—Pickup winding
N_S—Saturating winding
V_1, V_2—Each 10-6AS7 dual triodes (20 sections) in parallel

core is subjected to a magnetomotive force given by the number of ampere-turns applied. This means that if the radial build-up of the core is in the ratio of 2–1, the magnetic field intensity H in ampere-turns/meter at the ID (inside diameter) is always twice that at the OD (outside diameter). The surfaces that are loci of points of equal magnetic intensity are cylindrical surfaces concentric with the geometry of the core. For this reason, the basic postulate underlying all of these studies is that when the flux in the core is being reversed, from one saturation state to another, for example, the reversal tends to begin at the inner diameter of the core since the magnetic intensity is always highest there. Thus, in the course of a simple flux reversal, one would expect to find the inner portion of the core with its flux reversed, and the outer portion of the core with its flux not yet reversed. Between these two portions of the core lies what is called a "transition region."[3]

The question of the structure and the mode of motion of the transition region has been the subject of some discussion among workers in this field, and is still the object of research. Nevertheless, some reasonable postulate must be assumed in order to give direction to research. The concept presently favored by the authors is that the transition region extends through several wraps of tape, giving the cross-sectional appearance of a thin cylindrical section of the core which increases in radius as the magnetization reverses. Because of the space between wraps, it is assumed that this radial motion is equivalent to a linear motion of the transition region along the tape. In cross section, it appears as a spiral motion corresponding to the spiral winding of the tape. The rate of the linear motion along the tape is called the "transition-region velocity." It is a sort of "phase velocity" produced by the nucleation and growth of magnetic domains and the co-ordinated motion and annihilation of domain walls. All stages of flux reversal occur simultaneously, but are distributed spatially within the transition region, progressing from the initial stages in its outer portion to the final stages in the inner portion. In detail, domain nucleation and growth should be recognized as a statistical problem. When the grains are highly oriented or the crystal anisotropy is low and the material is properly prepared, walls begin to grow about most of the nucleation sites at fields distributed narrowly about some mean value. Macroscopically, this is commonly recognized as a threshold field, but it should not be interpreted as the threshold field defined in experiments on wires under tension or other experiments on single crystals.[4–6] Furthermore, in most flux reversal problems the energy required to reverse the magnetization of a given volume of material depends upon the time taken to reverse it and the structure of the domain walls involved.

Obviously, according to this picture, domain-wall velocity and transition-region velocity are not synonymous. The physical model of flux reversal is intentionally abbreviated in this paper because the major subject under discussion is the measurements performed during slow flux reversals.

A second point of view fundamental to this work is that there should be, in addition to the physical model of flux reversal, a mathematical model helpful in defining the parameters descriptive of the macroscopic behavior of the core. There is much merit in using the simple linear force-action equation (so long as care is taken not to abuse it in its application)

$$e_c = K(H - H_T), \text{ for } H > H_T \qquad (1)$$

$$e_c = 0 \text{ for } H \le H_T \qquad (1A)$$

where e_c is the induced voltage per turn in a winding, and H is the applied field at the center of the transition region. Here, attention is given to parameters K and H_T, and both must be regarded generally as functions of flux, rate of flux change, and past history of the core. The second basic equation, which results from the cylindrical symmetry of the core, gives the field H at the center of the transition region:

$$H = \frac{N_D i_x}{l} \qquad (2)$$

where i_x is the exciting current, and N_D the number of turns in the drive winding in which i_x flows. l is the magnetic path length, the circumference of the cylinder concentric with the core geometry and passing through the center of the transition region.

In using these two equations, it is important to recognize that the magnetic materials in the experiments reported in this paper were not under tension as they were in the experiments of Sixtus and Tonks,[4] Dijkstra and Snoek,[5] and Cole.[6] It must not be inferred from the similarity of equation 1 and the Sixtus-Tonks equation that equation 1 is descriptive of a fixed domain-wall configura-

Table I. Data on Cores and Windings

Characteristic	Core Magnetics No. 50041-2A	 Magnetics No. 50017-2A	 Special No. 1057	 Magnetics No. 50026-2D	 Magnetics No. 50026-2K
Material	Orthonol (grain-oriented) 50-50 Ni-Fe)	Orthonol	Special*	Square HyMu 80 (4Mo-79Ni-Fe)	Magnesil (grain- oriented 3%Si-Fe)
Tape thickness, mils	2	2	2.1	2	2
Tape width, inches	1	1/2	1/4	3/8	3/8
Core ID, inches	2	2	1.625	1	1
Core OD, inches	2.5	2.5	2	1.5	1.5
Mean magnetic path length, cm	18	17.9	14.4	9.84	9.84
Core cross-section area, cm²	1.37	0.685	0.26	0.514	0.514
Length of tape (estimated), meters	19.4	18.9	11.4	10.3	10.3
Number of wraps (estimated)	106	106	77	109	109
Core box	Phenolic	Phenolic	Phenolic	Phenolic	Phenolic
Approximate year of manufacture	1955	1959	1954	1959	1959
Feedback winding N_F	5,200 turns no. 40 wire	10,000 turns no. 40 wire	16,000 turns no. 40 wire	1,000 turns no. 40 wire	1,600 turns no. 40 wire†
Pickup winding N_P	4,700 turns no. 40 wire	9,400 turns no. 40 wire	8,300 turns no. 40 wire	9.800 turns no. 42 wire	11,700 turns no. 42 wire
Monitoring winding N_M	none	none	none	22,000 turns no. 42 wire†	8,860 turns no. 42 wire†
Drive winding N_D	18 turns no. 14 wire	18 turns no. 14 wire	15 turns no. 14 wire	10 turns no. 22 wire	10 turns no. 22 wire
Saturating winding N_S	18 turns no. 14 wire	18 turns no. 14 wire	15 turns no. 14 wire	10 turns no. 14 wire	10 turns no. 14 wire

* The special no. 1057 core is made of 50-50 Ni-Fe grain-oriented material, specially processed for low coercive force, and was furnished by M. F. Littman of Armco Steel Corporation.

† On the 50026-2K core, N_F was not used; N_M was used for both monitoring and feedback. On the 50026-2D core, N_F was used only for feedback; the oscilloscope and the Sanborn were connected to N_M.

tion moving along the tape at a velocity governed by the applied field.

Equations 1 and 2 may be combined into the following equivalent forms, which express the measurable electrical quantities explicitly.

$$i_z = \frac{l}{N_D}\left(\frac{e_c}{K}+H_T\right), \text{ for } e_c>0 \qquad (3)$$

$$e_c = K\left(\frac{N_D}{l}\, i_z - H_T\right), \text{ for } e_c>0 \qquad (4)$$

Equation 1 leads to the interpretation of H_T as the threshold field which must be exceeded by H in order to produce a change of flux in the core, indicated by a nonzero value of e_c. It describes the macroscopically measured threshold field corresponding to a particular domain-wall configuration in a particular part of the core. When the transition-region velocity is other than zero, the domain-wall structure is not in static equilibrium, and the question is how one then defines a threshold field for these moving walls. This leads to a third postulate, which is essential to the macroscopic mathematical model of the process.

The third postulate fits in with the macroscopic equations 1 and 2 and permits us to deduce the quantities K and H_T from measurements rather than from microscopic physical considerations. It is first postulated that over small ranges of flux and e_c, K and H_T are not functions of e_c or H, but are functions of history and flux alone. Therefore, from equations 1–4 it can be seen that at any value of flux, and for some prescribed history,

$$K = \frac{\partial e_c}{\partial H} \qquad (5)$$

a mobility parameter for the transition region. Equation 1 can be written

$$H_T = H - \frac{e_c}{K} \qquad (6)$$

in which the function H_T is known if K is also known at the same value of flux. It is assumed that the induced voltage arising from the rate of flux change in the core can be controlled experimentally. Over a set of experiments performed within a small range of values of e_c it is further possible to measure an average of the parameter K which can be defined as a constant, $\bar{K}$, for the core, and which is merely a measure of the average loop widening that occurs with increasing rate of flux change. Since the differential of energy required to produce a differential of flux, $d\phi$, in the magnetic structure is

$$du = (lH)\, d\phi \qquad (7)$$

the total measured magnetomotive force is the flux rate of change of energy put into the core. This can be represented in terms of equation 6 as

$$du = \left(H_T + \frac{e_c}{\bar{K}}\right) l\, d\phi \qquad (8)$$

The constant term $e_c/\bar{K}$ is a measure of the average energy that must be added to maintain the desired rate of flux change, and the variational term H_T is a parameter that measures an energy variation presumed to be associated in some way with the variation of metallurgical properties along the tape. H_T is a nonunique function of flux, whose properties change with previous history and rate of flux change.

The physical and mathematical model of low-speed flux reversals, contained in the previously mentioned three postulates, formed the basis on which many experiments, a few of which are described in this paper, have been performed on many cores of different materials and dimensions. It seems obvious that in order to emphasize the effects of metallurgical variations throughout the core and to study do-

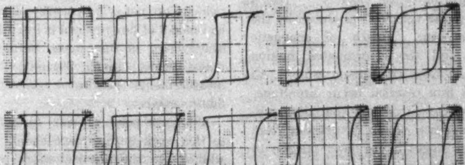


Fig. 2. "Sine flux" hysteresis loops at 60 and 400 cycles. (Loops for four of the five cores are shown. By accident, loops for the 50017-2A core were not photographed, so the fifth set of loops was taken on another core having similar characteristics.) Upper row: 60-cycle loops; lower row: 400-cycle loops. Scales are not calibrated. The vertical scale is changed for each loop. The horizontal scale is the same for both loops on a given core, but changes from core to core. The maximum saturating H for each loop is adjusted to 2 H_o, where H_o is taken as half the loop width at B=0. The excitation source impedance is approximately 0.1 ohm, including the resistance of the 10-turn exciting winding

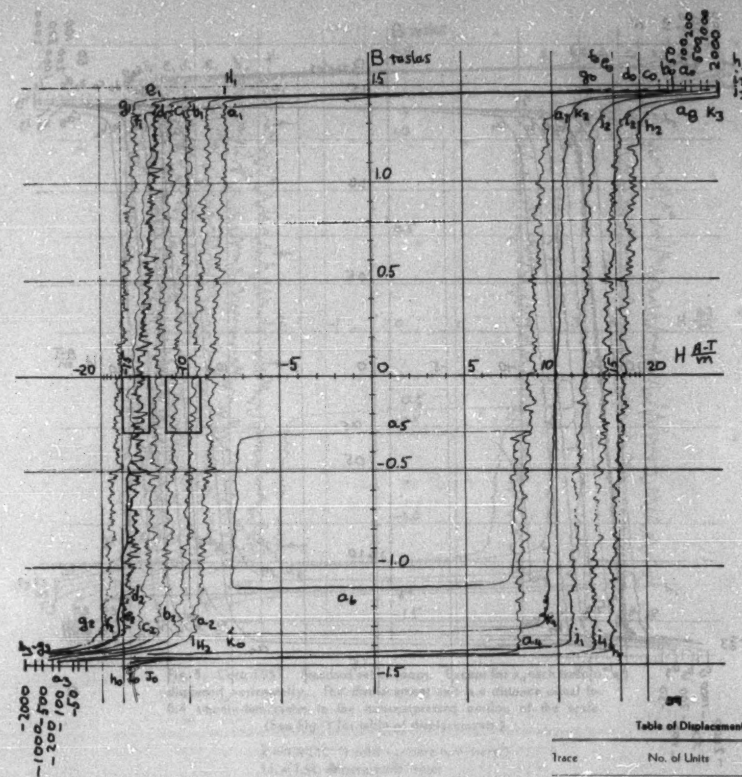


Fig. 3. **Core 50041-2A. Standard set of loops. Except for a, each trace is displaced horizontally. The displacement unit is a distance equal to: 1.0 ampere-turn/meter in the noncompressed portion of the scale** $\bar{K}=1.5\ (10^{-6})$ volts/(ampere-turn/meter); $H_c=8.8$ ampere-turns/meter; ID/OD=0.80 and $H_1/H_2=0.79$ The portions of traces b, c, e, and f enclosed in rectangles were recorded simultaneously on the Sanborn, and are shown in Fig. 8. In Figs. 3–7 the original set of loops is 2.5 feet square. Also, 1 tesla=1 weber per square meter=10 kilogauss

Table of Displacements

Trace	No. of Units	Direction
a	none	
b	1	to the left
c	2	to the left
d	3	to the left
e	4	to the left
f	4	to the left
g	5	to the left
h	5	to the right
i	4	to the right
j	3	to the right
k	2	to the right

main-wall motion through external electrical measurements, the rate of flux change and thus e_v/K (in equation 6, the average velocity-dependent portion of the total field) should be kept small, so that H is an approximate measure of H_T. Also, as the rate of flux change is made extremely small, the quantity $H \approx H_T$ approximates the threshold field of the static-equilibrium domain-wall configuration. When the behavior of the magnetic materials under these uncomplicated conditions has been thoroughly established, it should then be possible to increase the rate of flux reversal, and by noting changes in the measured quantities to infer the changes that take place in the structure of the transition region. Instrumentation has been arranged to permit establishment of the average rate of flux change as the independent variable, and measurement of the resulting magnetomotive force. (The magnetomotive force is plotted as the magnetic intensity at the mean path length of the core, to facilitate comparisons between cores of different materials and dimensions.)

A standard set of measurements has been made on each of the cores. The standard is the velocity of the transition region as it passes midway between the ID and OD of the core. The corresponding value of average rate of flux change, e_c, is maintained throughout the flux reversal, giving rise to a change in transition-region velocity of 20%, for example, in a core with an ID/OD ratio of 0.8. The induced voltage per turn, e_c, is thus a function of core dimensions and tape thickness, and is in the vicinity of 1 microvolt/turn at the standard

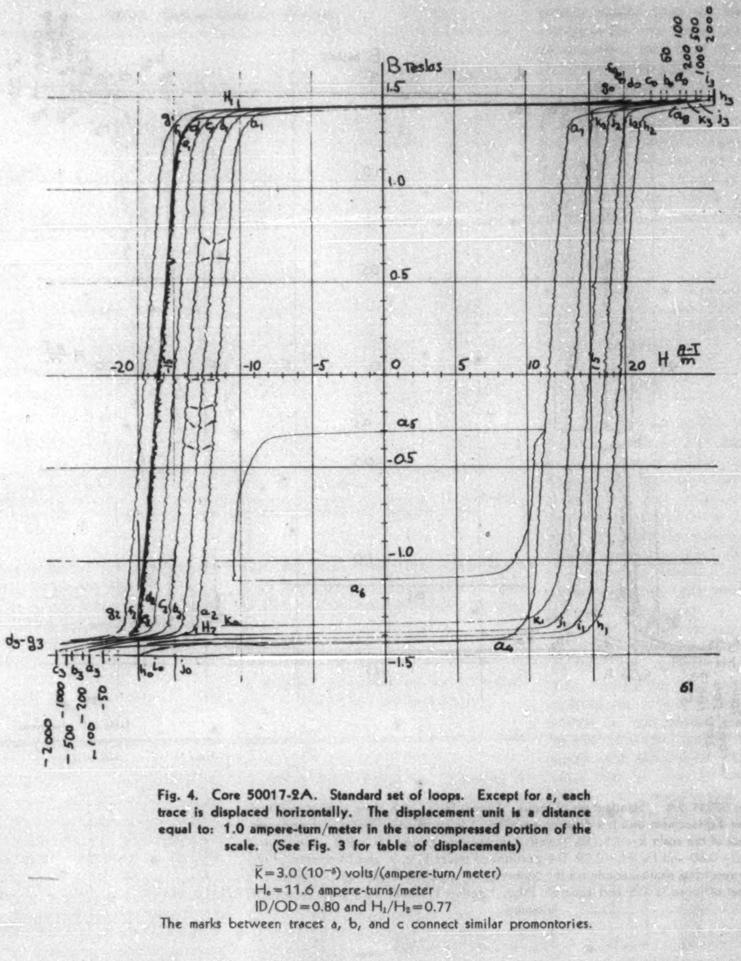


Fig. 4. Core 50017-2A. Standard set of loops. Except for a, each trace is displaced horizontally. The displacement unit is a distance equal to: 1.0 ampere-turn/meter in the noncompressed portion of the scale. (See Fig. 3 for table of displacements)

$\bar{K}=3.0\,(10^{-5})$ volts/(ampere-turn/meter)
$H_c=11.6$ ampere-turns/meter
$ID/OD=0.80$ and $H_1/H_2=0.77$
The marks between traces a, b, and c connect similar promontories.

velocity of 10 cm/sec (centimeters/second). Measurements of $\bar{K}$ for Orthonol place it in the vicinity of 10^{-5} volts/ampere-turn/meter. Thus, at the standard velocity, the term $e_c/\bar{K}$ has a value of about 0.1, which is in the order of 1% of H_F in equation 3. Transition-region velocities an order of magnitude below this value are very difficult to control. Many months of instrumentation refinement were required to make possible the experiments which are recorded in this paper.

Instrumentation

A block diagram of the circuitry used to perform the control, measurement, and calibration functions is shown in Fig. 1. The core driver, a specially designed d-c coupled power amplifier with low output impedance, derives its input from a subtracting circuit that compares the input voltage with the induced voltage on the feedback winding, N_F, of the core. The feedback system acts to maintain the induced voltage e_c within a few per cent of its desired value. Also, in order to trace hysteresis loops at a constant rate of change of flux and yet to limit the saturation magnetic intensity reached, an adjustable-threshold current-limiting circuit is incorporated into the feedback system, designed to override the voltage feedback when a desired current is reached, thus converting the input to current control. The excitation to the core is supplied by a low-voltage drive winding, N_D, having a few turns of heavy-gage wire. A separate saturating wind-

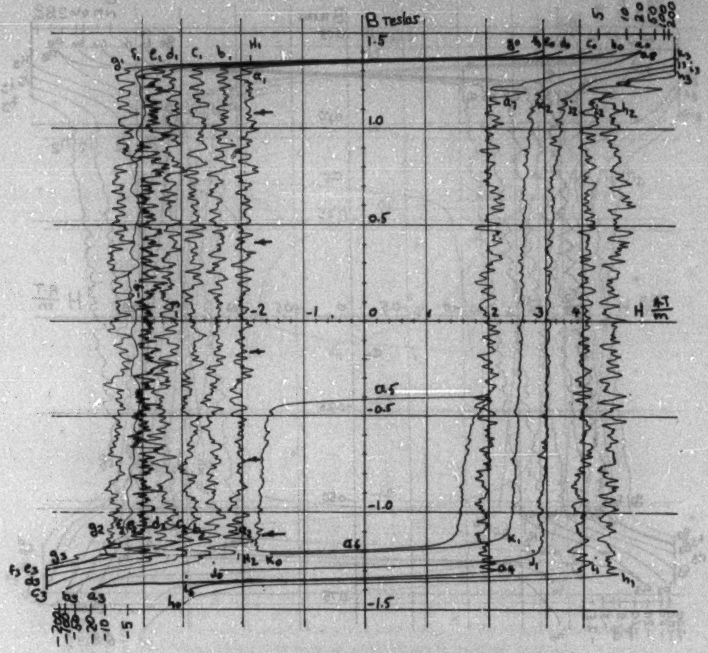


Fig. 5. Core 1057. Standard set of loops. Except for a, each trace is displaced horizontally. The displacement unit is a distance equal to: 0.4 ampere-turn/meter in the noncompressed portion of the scale. (See Fig. 3 for table of displacements.)

$\bar{K} = 1.8\ (10^{-6})$ volts/(ampere-turn/meter)
$H_e = 1.96$ ampere-turns/meter
ID/OD $= 0.81$ and $H_1/H_2 = 0.89$

The arrows show flux levels corresponding to points in the core at which there is a break in the tape

ing, N_S, is used in conjunction with a d-c power line to obtain high degrees of saturation not obtainable with the electronic amplifier. A nonlinear scale-compression circuit is used in the H-signal amplifier to bring the high saturation region onto the chart with the major loop. This amplifier, as well as others necessary for the plotter, and all of the calibrating circuits are components of a Reeves Analog Computer. Noise has been reduced so that it is less than the line width found on the H axis of the plotter.

The pickup winding, N_P, is used to develop the voltage which the integrator converts to a flux-density signal. The integrator employs a KinTel low-drift d-c amplifier with feedback through a capacitor. It has some drift, most of which can be compensated electronically. It is reset on each return of the core to positive saturation. In the measurements the drift is negligible in most cases. The entire hysteresis loop is recorded on an Electronics Associates X-Y plotter, $2^1/_2$ feet square, which has a bandwidth of about 1 cps (cycle per second). To obtain an expanded recording of faster magnetomotive force variations, a Sanborn strip-chart recorder is used. This has a bandwidth of approximately 60 cps. Comparisons with the oscilloscope, which has a bandwidth of about 500 kc, have shown that all significant variations can be seen on the Sanborn recordings. In summary, it should be noted that the arrangement of instrumentation is a study in itself, and is far from being a routine task.

Standard Measurements

Measurements presented include some of those taken on five of the cores tested; they are based on the plots of hysteresis loops using the instrumentation decribed. Four of the cores were supplied from stock by Magnetics, Inc., and the fifth, one of a set studied in a magnetic amplifier, (eighth core from the bottom in Table I of reference 7), was furnished by M. F. Littman of Armco Steel Corporation. The characteristics of these five cores and their windings are listed in Table I. The cores were chosen to show the variation of properties between different types of materials and between different cores of the same type of material.

Before any windings were placed on

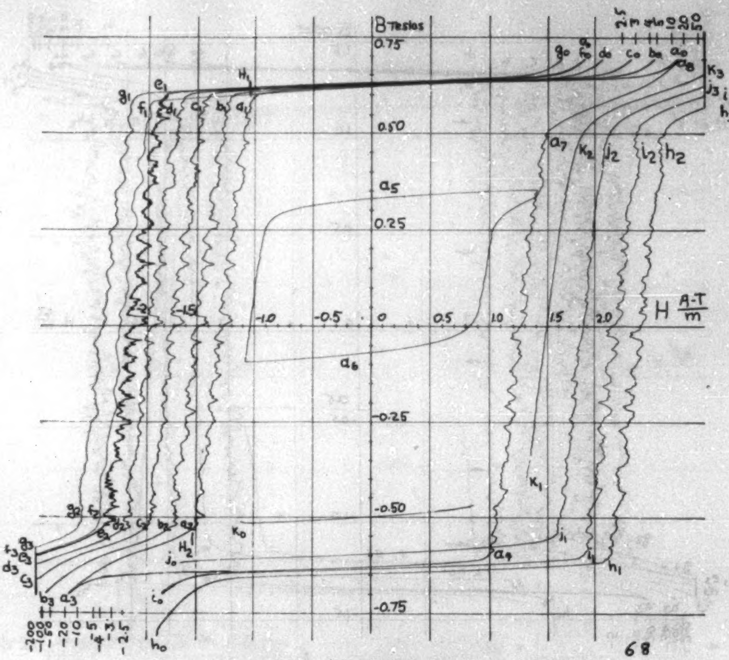


Fig. 6. Core 50026-2D. Standard set of loops. Except for a, each trace is displaced horizontally. The displacement unit is a distance equal to: 0.2 ampere-turn/meter in the noncompressed portion of the scale. (See Fig. 3 for table of displacements)

$\bar{K} = 5.0\ (10^{-6})$ volts/(ampere-turn/meter)
$H_0 = 1.27$ ampere-turns/meter
$ID/OD = 0.67$ and $H_1/H_2 = 0.68$

the cores, 60- and 400-cycle sinusoidal-flux hysteresis loops were photographed as shown in Fig. 2. Then each core was wound and connected to the equipment for a preliminary measurement of coercive force H_c. The saturation limits in the current-feedback circuit were set to $\pm 10\ H_c$. With this information and the design figures for the windings, the co-ordinates of the chart, including the built-in nonlinear H-scale, were calibrated and ruled. The core was then cycled between $\pm 10\ H_c$ ten times, and the sequence of experiments described subsequently was performed. The results for each core are shown in Figs. 3–7.

First, a major hysteresis loop with a small minor loop included was recorded in its proper position on the co-ordinate axes. For this loop the input voltage was adjusted to produce -10 millivolts on the feedback winding, which corresponds to the standard transition-region velocity of 10 cm/sec. The induced voltage was maintained at this value while the curve was traced from a_0 to a_2 The induced voltage was then set to $+10$ millivolts while the trace was recorded to a_5. There the induced voltage was set to -10 millivolts tracing to a_6, where it was set to $+10$ millivolts again, and maintained while the minor loop closed and the flux continued its reversal to a_8. In order to show the extent to which the variations in H are reproduced on successive loop sides, two more traces were made of the left side of the loop under experimental conditions as close as possible to those of loop a. For trace b, the pen was displaced one unit to the left, and the induced voltage was set at -10 millivolts, while the trace b_2–b_3 was recorded. The return to positive saturation was made without tracing, the pen diplaced a second unit to the left, and trace c recorded under the same conditions. In order to see whether there was any obvious polarization of the core, the return to positive saturation was recorded by inverting the plotter signals with a pair of unity-gain computer amplifiers, displacing the pen a third unit to the left and recording the right side of the loop on the left side of the chart, trace d, under the same experimental conditions.

Next, some recordings were made to evaluate the effects of changes in transition-region velocity. The plotter was restored to its original, noninverted condition, and the pen was displaced a fourth unit to the left. Two traces were made from the same saturation levels as be-

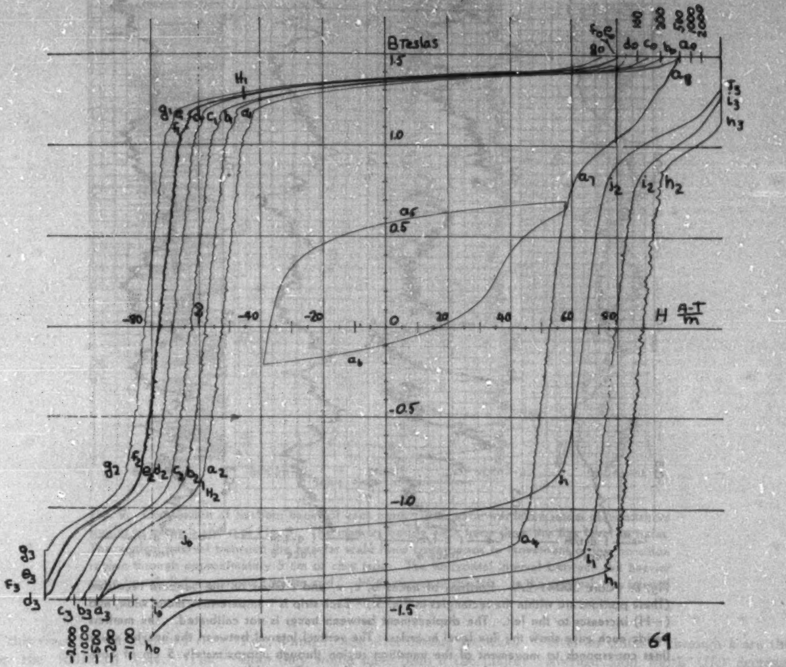


Fig. 7. Core 50026-2K. Standard set of loops. Except for a, each trace is displaced horizontally. The displacement unit is a distance equal to: 5.0 ampere-turns/meter in the noncompressed portion of the scale. (See Fig. 3 for table of displacements; note: no trace k for this figure)

$\bar{K}=4.0\ (10^{-6})$ volts/(ampere-turn/meter)
$H_c=53$ ampere-turns/meter
ID/OD$=0.67$ and $H_1/H_2=0.75$

fore, without further pen displacement. The first, trace *e*, corresponds to an induced voltage of 2.5 millivolts and the second, trace *f*, corresponds to 50 millivolts. Thus, the ratio of speeds is 20–1 in these traces. In each case the core was returned to positive saturation without recording a trace.

Traces *h* through *k* show the effects of the degree of saturation reached just prior to tracing the loop side. With the current-feedback threshold set to limit the field to $+10H_c$, each of the four traces was begun at positive saturation by adding a current in N_s to bring the total field to a reference level of $+100\ H_c$. The four traces, *h*, *i*, *j*, and *k*, were taken with the pen displaced five, four, three, and two units to the right respectively. The traces were made after the flux had been brought from the positive reference $+100\ H_c$ to, respectively, $-100\ H_c$, $-10H_c$, $-2H_c$, and a point on the loop side just short of saturation.

Analysis of Results

Presentation of the experimental results in the familiar form of the hysteresis loop is useful for the perspective it lends, and for the purpose of rapid qualitative evaluation of the data. Quantitatively, however, the recordings lack detail in this form because of the response of the *X-Y* recorder, and because the scale is adjusted to get the entire loop on the chart. The Sanborn strip chart recorder was therefore used to collect simultaneous supplementary data with expanded *B* and *H* scales. Some of these recordings will be introduced in the following to assist in analyzing the results.

The character and amplitude of the variations in *H* differ widely for various materials and for different cores of the same material. A comprehensive analysis of the measurements is a monumental task which, in fact, is not yet complete. There are some outstanding characteristics, however, which deserve comment. For most cores, the recording of *H* seems to consist of a contour having coarse variations with superimposed fine detail. On successive experiments,

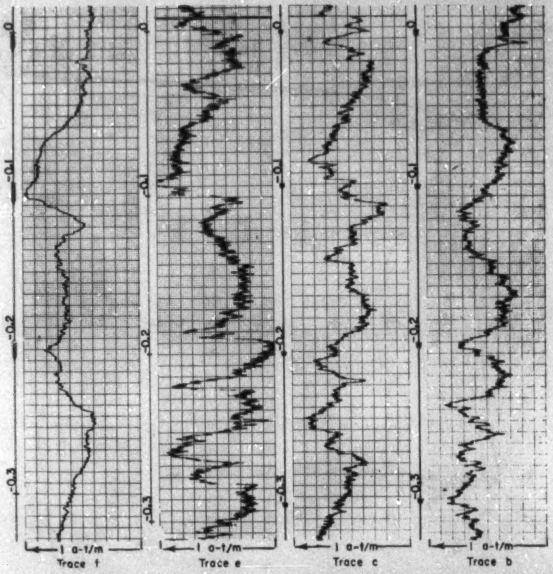


Fig. 8. Core 50041-2A. Portions of traces b, c, e, and f taken on the Sanborn recorder. (These portions are within the rectangles in Fig. 3.) Each strip is 1 ampere-turn/meter wide, and (−H) increases to the left. The displacement between traces is not calibrated. The markers beside each trace show the flux level in teslas. The vertical interval between the heavier scale lines corresponds to movement of the transition region through approximately 5 cm of core tape. In the original, each strip was 5 cm wide

such as traces *a–d*, it is found that the fine detail is not repeatable, but that in some measure the coarse variations are. The degree of reproducibility differs greatly from core to core, but on practically every core tested, there are enough reproducible promontories on a loop side to give it a distinctive, recognizable profile. Of the five cores included in this paper, Fig. 6 shows the most reproducibility, Fig. 7 the least. One characteristic frequently found is that repeatable features may recur at slightly different flux levels, as shown in Fig. 4. This substantiates the argument that the transition region shape and motion are not unique, and that it is not possible to make an exact correlation between flux level and position of the transition region, although there is a close relation between them. The spread in flux level of the reproducible promontories gives some idea of the deviation to be expected.

On the five loops shown, there is no more and no less correlation between the contours *a* and *b*, for example, than there is between *a* and *d*. Under these experimental conditions, therefore, suspicions that the core might have some permanent polarization are shown to be unfounded.

The difference between fast and slow loops is seen in traces *e* and *f*. The average displacement to the left, due to the velocity component of energy that must be added, differs from core to core. However, changes in plotter scale affect the apparent displacement; a better comparison is given by the measured values of $\bar{K}$ shown on each figure. There is a remarkable difference between these two traces in the amount of detail in the variations of H. The filtering effect of the X-Y plotter is most pronounced on the high-speed loop, but the Sanborn recordings, which miss no significant detail, show the same phenomenon. A typical set of these recordings, for the *50041-2A* core, is shown in Fig. 8. It can be seen that the fine detail is affected by velocity changes much more than the coarse variations. For the purpose of orientation with respect to the frequency of these loops, it should be noted that a standard loop, traces *a–d*, in which the transition region velocity is 10 cm/sec, corresponds to 1 cycle in approximately 6 minutes on the *50041-2A* core. The fast loop corresponds to 1 complete cycle in about 1.2 minutes, and the slow loop to 1 cycle in 24 minutes. The speed of the strip chart is changed with the transition region velocity so that a given length of chart corresponds to approximately the same length of core tape for all traces.

Fig. 9 shows sections of the strip charts for equal tape lengths in the five cores, taken with equal gains in the H-signal amplifier. Each is taken at standard velocity from standard saturation. These provide a better comparison between cores than the plotter charts, since the scale of the latter changes from core to core. The differences in the character and amount of coarse and fine detail are apparent.

Trace *g* in each of the standard sets was taken while measuring the average mobility parameter K. Fig. 10 shows a

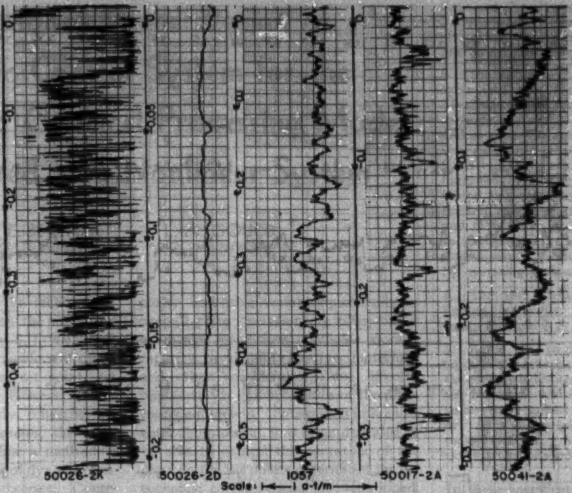


Fig. 9. Comparison of Sanborn traces of each core. Each trace was taken under circumstances similar to traces b and c of Fig. 8. The markers beside each trace show the flux level in teslas. The vertical interval between the heavier scale lines corresponds to movement of the transition region through approximately 5 cm of core tape. The horizontal interval between the heavier scale lines is 0.1 ampere-turn/meter and (−H) increases to the left. In the original, this interval was 0.5 cm

section of this trace on the Sanborn recorder for the *50017-2A* core. The induced voltage steps, checked with the oscilloscope, were very square. The overshoot shown in Fig. 10 is that of the recorder. The height of the voltage step divided by the average increment of H is the value of $\bar{K}$ shown for the core.

If the assumed model of flux reversal is correct in general, with no fast commitments to detail, the ratio of H_1 "and" H_2 (the beginning and end of the straight portion of trace a) should equal the (ID/OD) ratio of the core. The comparison is made on the illustrations, and the ratios are found to be in reasonably good agreement. (In Figs. 5 and 7 the sides of loop a are significantly steeper than predicted, but traces k and j, respectively, have nearly the predicted slope.) Also, the model predicts[3] the nature of the minor loop taken in the middle of trace a. The different ways in which the minor loops close for different cores, considered together with other parts of the data shown here and with still other experiments, should be helpful in clarifying the details of transition-region movement. It is particularly interesting to note, for example, that the HyMu *80* and Magnesil minor loops, Figs. 6 and 7, just after closing are wider than they would have been if the major loop had not been interrupted to take the minor loop.

Another interesting result of the measurements is the large size of the coarse variations on some cores. In Fig. 3, for example, some of the larger promontories cover a flux increment corresponding to about two wraps of tape, over which the measured H is 5% higher than in the immediately adjacent sections of the core. It is not quite clear how these should be interpreted, because gross physical discontinuities introduced purposely into a core do not produce such marked variations in measured H. For example, the core in Fig. 5 is made of six separate strips of tape wound in succession and overlapped where the joints occur. The flux levels corresponding to the locations of these joints are marked in the illustration. Whatever the variations due to these lapped joints are, they are certainly no larger in order of magnitude than those occurring elsewhere in the continuous tape, or they would appear prominently in the loop. The same lack of prominent effect has been found for 10-mil-diameter holes drilled in the tape.

The principal effects of prior saturation level on traces h through k are that the average loop width (H_c) decreases, as well as the amplitude of the coarse variations. Sanborn recordings show that the amount of fine detail is relatively unaffected. Some cores respond dramatically to saturation level changes, others very little. These results can be explained in terms of the physical model, thereby tending to confirm it. The coarse variations are presumed to represent changes in the domain wall arrangement, caused in some way by metallurgical irregularities much more extensive than simple crystal imperfections, which change the energy required to move the transition region. These areas of irregularity would act as centers of unreversed magnetization left behind the transition region, and thus as nuclei for reverse magnetization when the flux is again reversed. The presence of these nuclei makes the reversal smoother and easier. Most of them are destroyed, however, during the irreversible flux change in saturation.

Differences noted between types of materials follow from differences in such things as grain orientation, but even more interesting are the differences found in a given type of material. Figs. 3 and 4

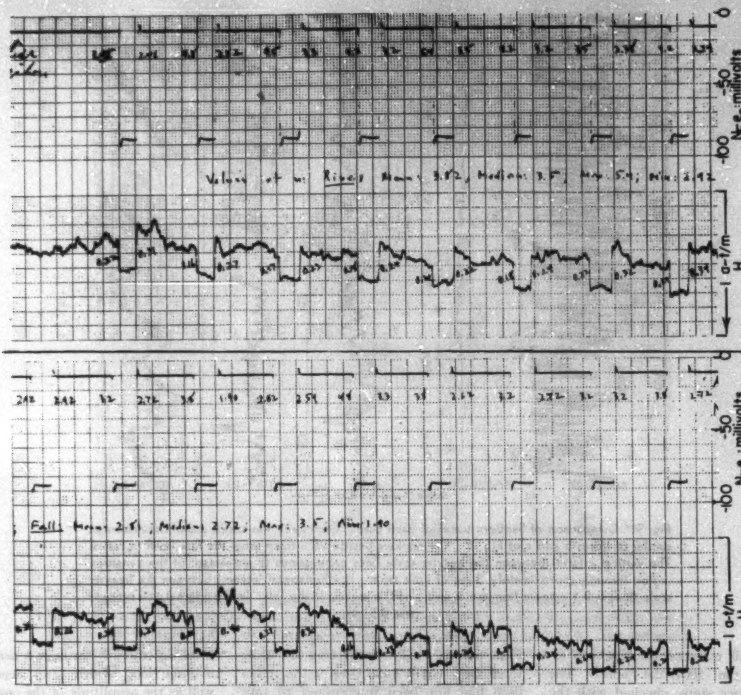


Fig. 10. Core 50017-2A. Step changes in e_c and I_x. Two successive portions of a single recording are shown. In each, the upper trace shows the voltage induced on the feedback winding. On the current trace, the numbers beside the steps give their height in ampere-turns/meter. The numbers beside the voltage steps show n, where K=(10^{-8})n. Shown between the traces are averages and extremes of n for the left-hand, or "rise" steps, and for the right-hand, or "fall" steps. Since e_c is negative, it "rises" downward and "falls" upward, and so does (−H). The Sanborn speed was 5 cm/sec. In the original, each trace was 5 cm high

typify the two rough classes found in Orthonol. It seems apparent from the extent of loop widening measured that whatever the nature of the irregularities which give rise to these variations in H, they are more finely and more uniformly distributed, with larger wall area, in the core of Fig. 4 than in that of Fig. 3.

Conclusions

The experiments reported in this paper tend to confirm the suggested physical model of slow flux reversal. But more important, the experiments resulted in a number of facts which must be explained by any physical model and which call for further experiments. The primary purpose of this paper is to show the details of measurements on tape-wound cores under low fields. A quantitative analysis of the constants K and H_T is summarized elsewhere.[8]

References

1. Applications of Nonlinear Magnetics, Herbert F. Storm. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, July 1958, pp. 380–88.

2. Magnetism in Tape-Wound Cores, R. C. Barker. *Ibid.*, vol. 79, Nov. 1960, pp. 482–503; also "Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication*, 1958.

3. Direct-Current Measurements on Tape-Wound Cores, R. C. Barker, R. M. Brownell. *Journal of Applied Physics*, New York, N. Y., vol. 31, 1960, pp. 243S-44S.

4. Propagation of Large Barkhausen Discontinuities, K. J. Sixtus, L. Tonks. *Physical Review*, New York, N. Y., vol. 43, 1933, pp. 931–40.

5. Propagation of Large Barkhausen Discontinuities in Fe-Ni Alloys, L. J. Dijkstra, J. L. Snoek. *Philips Research Reports*, Eindhoven, Netherlands, vol. 4, 1949, pp. 334–56.

6. Motion of Ferromagnetic Domain Walls in 48% Ni-Fe Tapes, R. W. Cole. *Journal of Applied Physics*, vol. 27, 1956, p. 1104.

7. Relation of D-C Magnetic Properties of Oriented 48-Per-Cent Nickel-Iron to Magnetic-Amplifier Performance, C. E. Ward, M. F. Littman. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 74, Sept. 1955, pp. 422–27.

8. Measurement of the Domain Wall Area—Mobility Product During Slow Flux Reversals, R. M. Brownell, R. C. Barker. *Journal of Applied Physics*, vol. 32, 1961, pp. 284S–85S.

Considerations in the Design of a Dynamic Control System for the Generalized Chemical Processing Model Considered as a Nonlinear System

A. E. BEECHER
ASSOCIATE MEMBER AIEE

L. A. GOULD
ASSOCIATE MEMBER AIEE

THE CONTROL of the reactor and associated plant units in the processing complex hypothesized by T. J. Williams[10] and specifically, the dynamic control of this reactor to a desired set point, are treated in this paper. However, while the problem particularly concerns the control of the given reactor, the main purpose here is to outline a method of approach to the problem of dynamic control of industrial processes. Consequently, short cuts peculiar to the reactor control problem are ignored. In general, the approach is as follows:

1. Assumptions are made concerning criteria of operation and behavior of units of the plant other than the reactor proper.
2. From these assumptions the calculus of variations is used to develop control laws or strategies for each control variable.
3. Since these control laws cannot, in general, be integrated analytically, the prospect of using an "on-line" computer to integrate these laws numerically and consequently control the reactor is then examined.

In addition, the following topics are treated in the appendixes:

1. Appendix I examines the mechanics of the application of variational calculus to the problem of obtaining the control laws for the processing complex considered in this paper.
2. Appendix II discusses the application of the technique to a simple system. This is included to help the reader in understanding the technique and its capabilities.

The Problem and the Approach

The determination of such quantities as desired plant throughput, desired product purity, and other variables pertaining to over-all plant operation are often based on market conditions, available storage capacity, availability of raw materials, and other considerations not directly related to the characteristics of the processing equipment. When such influences enter the problem, it is often desirable to be able to control the operation of the plant to the set points determined by the economic considerations just mentioned. This implies that a dynamic control system can be designed which is capable of operation over a wide range of set points. The Williams plant is an example of just such a case. The operation of the plant is influenced by the cost of discarding by-product, a finite market for the main product, and occasionally an inability to handle cooling problems caused by excessive cooling water temperatures. Furthermore, various control variables and internal quantities must be limited to reasonable ranges of operation. Finally, any control system that is devised for the Williams plant should operate well regardless of where, in terms of set point, the plant is to operate.

The specifications for the control system can be developed from the following assumptions:

1. It is assumed that the desired weight fraction of recoverable product, denoted by P_0, is determined independent of the controller. This constitutes a set point which will be treated as an undetermined constant. It is assumed that P_0 is fixed by the operator and the control system is required to hold the production as close to this value as possible.
2. The temperature of the reactor contents is assumed to be determined independent of the controller. This is another set point and will be handled in the same manner as rate of product manufacture with the exception that variations of reactor temperature of ±10 F (degrees Fahrenheit) will be tolerated. This specification has been imposed because the manufacture of product and by-products is influenced by temperature. The tolerance serves as a limit for temperature change during transient conditions and also allows for the possibility that the temperature specification and the rate of product manufacture specification may not be entirely compatible.
3. The feed rate of reactant B and the recycle feed rate cannot become negative. Also, these feed rates must be limited to some finite value. Because of limitations due to pipe size and pumping capabilities, it is clear that these rates cannot have infinite ranges. It is assumed that a maximum feed rate of reactant B of 2 F_1 ft^3/hr (cubic feet per hour) is acceptable. Thus the feed rate of reactant B lies between zero and this value. F_1 is a variable available to the operator as a set point if necessary. Consequently, the limitation can be altered by operating personnel.
4. The feed rate of the recycle stream is limited (for similar reasons) to a value between zero and 2 F_1 ft^3/hr.
5. Again, due to physical limitations, the amount of heat supplied to or removed from the reactor is limited. Heat removed from the reactor is assumed positive and is assumed to be limited to the range $\pm H_0$ Btu/hr. This symmetrical limitation is chosen for convenience. A nonsymmetrical limitation can also be used if desired.

In these specifications nothing has been said of F_R, the reactor effluent flow rate. Although important in determining rate of manufacture, it is desired that this quantity have an average value but not necessarily a rigidly limited range. Also, it has been tacitly assumed that the other units in the plant are operating satisfactorily.

The five specifications listed may be written in the form of an integral criterion of plant operation. The control system design that results from the minimization of the integral criterion will thus be optimum within the limitations of the specifications. An integral form is chosen because it weighs errors regardless of when they occur and recognizes small long-term errors as well as larger errors which may not last very long.

Specification 1 concerns a quantity to be regulated precisely. A suitable penalty function for the criterion of plant operation is $K_p(P_R-0.1\ E_R-P_0)^m$. K_p is a weighting multiplier and m is an even integer. The use of the exponent m allows a nonlinear weighting as the errors become more pronounced; this exponent must be even to insure that the penalty function never becomes negative. Penalty functions may be developed for the other specifications in a similar manner. When variables are tolerated within a given range, the penalty function must express this. If the exponent used in the previous penalty function is allowed to become very large, then the penalty function increases very quickly as the value of $(P_R-0.1\ E_R-$

Paper 60-980, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE-AIChe-IRE-ISA-ASME Joint Automatic Control Conference, Cambridge, Mass., September 6–9, 1960, and re-presented for discussion only at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted May 5, 1960; made available for printing December 22, 1960.

A. E. BEECHER and L. A. GOULD are with the Massachusetts Institute of Technology, Cambridge, Mass.

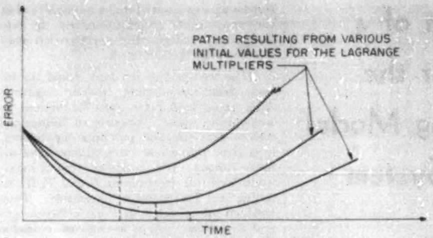


Fig. 1. Process of selecting t_1 as the point when error is a minimum

P_0) becomes greater than 1 or less than −1. Otherwise, the function is small. This provides the necessary form of penalty function for expressing a willingness to tolerate restricted change of the value of a variable.

The integral criterion of plant operation may now be written using the penalty function concept (see reference 10 for nomenclature not previously explained):

$$I=\int_0^{t_1}\Bigg\{(P_R-0.1E_R-P_0)^m+\left(\frac{T_R-T_0}{10}\right)^n+\left(1-\frac{F_B}{F_1}\right)^r+\left(1-\frac{F_L}{F_2}\right)^r+\left(\frac{H_w}{H_0}\right)^r\Bigg\}dt \quad (1)$$

where the weighting multipliers have been set equal to unity. This does not cause any loss of generality since the first term is supposedly the only term having any value as long as the other variables stay within their limitations. If these variables exceed the limits, then the appropriate penalty function should increase rapidly without the benefit of a multiplicative constant. Thus if r is a large even integer, then I is not greatly affected by control variables F_B, F_L, and H_w as long as they remain within their specified bounds. Consequently, by making m a smaller even integer, I becomes sensitive to product composition unless other variables (T_R, T_B, F_L, H_w) attempt to exceed their specified operating ranges. The problem thus is to minimize I over the time interval $t=0$ to $t=t_1\rightarrow\infty$ subject to boundary conditions and constraints yet to be imposed. If I is a minimum, the control system is optimum.

As stated above, I is to be minimized for a specific set of boundary conditions. These boundary conditions serve to limit the solution to the problem of interest (see Appendix II). At any instant of time the control system is faced with the problem of determining the best course of action to hold the plant at, or return it to, a desired group of settings. In either case it is said that "the control system is to *proceed* to this desired group of settings." This group is the desired steady-state operation of the plant. Therefore this is the boundary condition at time t_1. In other words, by specifying this steady-state condition as the boundary condition at time t_1, the control system that results will maneuver the plant to the desired steady-state condition. Any solution obtained must proceed to this steady-state condition. Similarly, the boundary condition at time $t=0$ is equal to the present state of the plant. This means the control system will solve the problem of how best to proceed from the present plant condition to the desired steady-state condition. To specify the boundary condition at $t=t_1$ requires a definition of the steady-state condition desired. This is accomplished in the form of a specification on desired average behavior (which will be approximately achieved if the control system always tends to return the system to the desired steady-state conditions):

1. The rate of manufacture of recoverable product shall, on the average, be as specified in the criterion of plant operation.
2. Reactor temperature shall, on the average, equal T_0.
3. Reactor effluent flow rate shall, on the average, equal a constant F_3. This constant is another set point and is treated as an undetermined constant available to the plant operator for setting.
4. The control variables shall be constant with respect to time. As will be seen later, the resultant control system will allow the derivatives of the control variables to be zero only when a steady-state condition is reached.

This group of specifications requires that a specific state of the system exist and that it shall not change with time. This is sufficient to specify the desired steady-state condition.

These specifications will now be put in mathematical form. Since they are to be used as boundary conditions at $t=t_1$ only, it is unnecessary to use an integral expression. Although any form of mathematical expression is acceptable, a quadratic form is used here for convenience; it is called the criterion of controller operation. This criterion, called Error, is used later in the actual operation of the control system.

$$\text{Error}=(P_R-0.1E_R-P_0)^2+(T_R-T_0)^2+\left(\frac{dF_B}{dt}\right)^2+\left(\frac{dF_L}{dt}\right)^2+\left(\frac{dH_w}{dt}\right)^2+(F_R-F_3)^2 \quad (2)$$

When Error is zero, the desired steady state exists. For the time being it will be sufficient to regard Error as a statement of the boundary condition at $t=t_1$ (see Fig. 1).

The constraint equations state the relationships between controlled and manipulated variables that are used in I, and thus are simply the differential equations that describe the dynamic behavior of the system. Certainly, if all the variables in I were independent, the solution would be to set them at their prescribed values. However, T_R, P_R, and E_R are the result of manipulating F_B, F_L, and H_w, and also the result of the uncontrolled changing of F_A, T_A, and T_B. Thus the optimum path is limited by the independence of the variables. As will

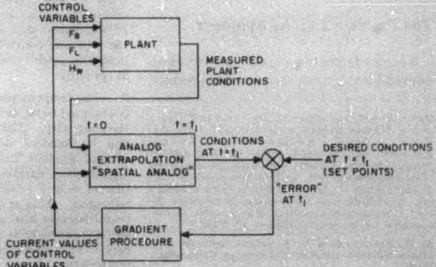

REACTOR CONTROL VARIABLES
F_B
F_L
H_W
PLANT
MEASURED PLANT CONDITIONS
t = 0
t = t_1
ANALOG EXTRAPOLATION "SPATIAL ANALOG"
CONDITIONS AT t = t_1
DESIRED CONDITIONS AT t = t_1 (SET POINTS)
"ERROR" AT t_1
GRADIENT PROCEDURE
CURRENT VALUES OF CONTROL VARIABLES

Fig. 2. Block diagram of the proposed Williams plant controller

be seen later, it is desirable to keep the number of constraint equations as small as possible since each constraint adds a degree of complexity to the resultant control system. For this reason, some variables have been eliminated where it is easy to do so in order to reduce the number of constraint equations needed. This is possible to some extent in most problems. (For example, F_R has been replaced by $F_A+F_B+F_L$, thus eliminating the variable F_R.)

There are 11 constraint equations. The first seven are derived from the reactor equations as specified by Williams:

$$\frac{dA_R}{dt}=0.01078[F_A+F_LA_L-A_R(F_A+F_B+F_L)]-5.98\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)A_RB_R \tag{3}$$

$$\frac{dB_R}{dt}=0.01078[F_B+F_LB_L-B_R(F_A+F_B+F_L)]-5.98\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)A_RB_R-2.6\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R \tag{4}$$

$$\frac{dC_R}{dt}=0.01078[F_LC_L-C_R(F_A+F_B+F_L)]+11.96\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)A_RB_R-5.2\,10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R-9.63\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)C_RP_R \tag{5}$$

$$\frac{dE_R}{dt}=0.01078[F_LE_L-E_R(F_A+F_B+F_L)]+5.2\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R \tag{6}$$

$$\frac{dG_R}{dt}=0.01078G_R(F_A+F_B+F_L)+14.45\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)C_RP_R \tag{7}$$

$$\frac{dP_R}{dt}=0.01078[0.1E_LF_L-P_R(F_A+F_B+F_L)]+2.6\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R-4.82\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)C_RP_R \tag{8}$$

$$\frac{dT_R}{dt}=0.01078[F_L(T_L-T_R)+F_A(T_A-T_R)+F_B(T_B-T_R)]-H_w+3,730\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)A_RB_R+975\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R+5,160\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)C_RP_R \tag{9}$$

The other four constraint equations are obtained by simplifying the remainder of the plant on the assumption it is operating properly. The rest of the plant is approximated as a perfectly mixed tank by merely computing the time lag in each unit and considering the total time lag to be the sum of the individual time lags. This procedure yields a set of equations which are used as additional constraint equations. They are not exact, but the dynamic characteristic that has been developed is approximately correct with regard to approximate time lag, steady-state behavior, and "direction of response." (The approximate system, when disturbed, moves in the same direction as the approximated system.) The resultant constraint equations are

$$\frac{dA_L}{dt}=2.665A_R-2.425A_L \tag{10}$$

$$\frac{dB_L}{dt}=2.665B_R-2.425B_L \tag{11}$$

$$\frac{dC_L}{dt}=2.665C_R-2.425C_L \tag{12}$$

$$\frac{dE_L}{dt}=2.665E_R-2.425E_L \tag{13}$$

Development of Control Laws

The problem now is how to manipulate F_B, F_L, and H_w to minimize I subject to the prescribed boundary conditions and constraint equations. This is a classical problem of variational calculus. The details of solution are considered in Appendix I. It is sufficient to state here that a time-dependent Lagrange multiplier, λ_i, arises from each constraint equation. The resultant control laws are the equations for F_B, F_L, and H_w in terms of these Lagrange multipliers and a set of differential equations whose solutions yield the multipliers as functions of time.

$$\frac{d\lambda_1}{dt}=0.01078\lambda_1F_R-2.665\lambda_8+5.98\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)B_R[\lambda_1+\lambda_2-2\lambda_3-625\lambda_7] \tag{14}$$

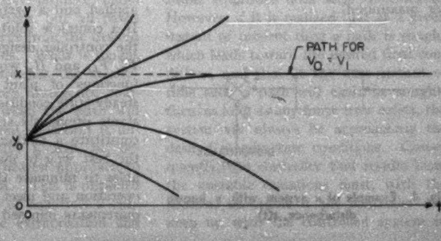


Fig. 3. A few trajectories the system may follow with various initial conditions

$$\frac{d\lambda_2}{dt}=0.01078\lambda_2F_R-2.665\lambda_9+5.98\times10^{9}\left(e^{-\frac{12,000}{T_R}}\right)A_R(\lambda_1+\lambda_2-2\lambda_3-625\lambda_7)+2.6\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)C_R(\lambda_2+2\lambda_3-2\lambda_4-\lambda_6-375\lambda_7) \tag{15}$$

$$\frac{d\lambda_3}{dt}=0.01078\lambda_3F_R-2.665\lambda_{10}+2.6\times10^{12}\left(e^{-\frac{15,000}{T_R}}\right)B_R(\lambda_2+2\lambda_3-2\lambda_4-\lambda_6-375\lambda_7)+4.82\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)\times P_R(2\lambda_3-3\lambda_5+\lambda_6-1,070\lambda_7) \tag{16}$$

$$\frac{d\lambda_4}{dt}=0.01078\lambda_4F_R-2.665\lambda_{11}+0.1m(P_R-0.1E_R-P_0)^{m-1} \tag{17}$$

$$\frac{d\lambda_5}{dt}=0.01078\lambda_5F_R \tag{18}$$

$$\frac{d\lambda_6}{dt}=0.01078\lambda_6F_R-m(P_R-0.1E_R-P_0)^{m-1}+4.82\times10^{15}\left(e^{-\frac{20,000}{T_R}}\right)\times C_R(2\lambda_3-3\lambda_5+\lambda_6-1,070\lambda_7) \tag{19}$$

$$\frac{d\lambda_7}{dt}=0.01078\lambda_7F_R-0.1n\left(\frac{T_R-T_0}{10}\right)^{n-1}+\frac{1}{T_R^2}\left\{7.17\times10^{13}\left(e^{-\frac{12,000}{T_R}}\right)A_RB_R\times(\lambda_1+\lambda_2-2\lambda_3-625\lambda_7)+3.9\times10^{16}\times\left(e^{-\frac{15,000}{T_R}}\right)B_RC_R(\lambda_2+2\lambda_3-2\lambda_4-\lambda_6-375\lambda_7)+9.64\times10^{19}\left(e^{-\frac{20,000}{T_R}}\right)\times C_RP_R(2\lambda_3-3\lambda_5+\lambda_6-1,070\lambda_7)\right\} \tag{20}$$

$$\frac{d\lambda_8}{dt}=2.425\lambda_8-0.01078\lambda_1F_L \tag{21}$$

$$\frac{d\lambda_9}{dt}=2.425\lambda_9-0.01078\lambda_2F_L \tag{22}$$

$$\frac{d\lambda_{10}}{dt}=2.425\lambda_{10}-0.01078\lambda_3F_L \tag{23}$$

$$\frac{d\lambda_{11}}{dt}=2.425\lambda_{11}-0.01078F_L(\lambda_4+0.1\lambda_6) \tag{24}$$

and

$$H_w = -H_6\left\{\frac{\lambda_7 H_6}{r}\right\}^{\frac{1}{r-1}} \quad (25)$$

$$F_L = F_2 + F_2\left\{\frac{0.01078F_2}{r}\Big[(A_R - A_L)\lambda_1 + (B_R - B_L)\lambda_2 + (C_R - C_L)\lambda_3 + (E_R - E_L)\lambda_4 + C_R\lambda_5 + (P_R - 0.1E_L)\lambda_6 + \lambda_7(T_R - T_L)\Big]\right\}^{\frac{1}{r-1}} \quad (26)$$

$$F_B = F_1 + F_1\left\{\frac{0.01078F_1}{r}\Big[\lambda_1 A_R + \lambda_2(B_R - 1) + \lambda_3 C_R + \lambda_4 E_R + \lambda_5 G_R + \lambda_6 P_R + \lambda_7(T_R - T_B)\Big]\right\}^{\frac{1}{r-1}} \quad (27)$$

It is generally impossible to proceed further analytically and find F_L, F_B, and H_w as functions of only plant conditions and desired boundary conditions at time t_1.

It is worth stopping at this point to consider what has been developed thus far. A dynamic control system was sought which would be capable of controlling the Williams reactor imbedded in the complete processing complex. A general discussion about the qualities of good dynamic control then followed. This resulted in a group of five specifications which the dynamic control system should meet. Each of these specifications was then cast in a mathematical form termed a penalty function. The integral of the sum of these penalty functions over a semi-infinite interval was the criterion of plant operation. This criterion was to be minimized to obtain the best possible operation. It is arbitrary in many respects and is intended *only* as an expression of desired dynamic behavior to be minimized or maximized for best performance. Any sort of penalty function is admissible as long as it is analytic and the operations of Appendix I are properly defined mathematically. This allows the use of a wide variety of functional forms. In addition, it is not necessary that the specifications be mutually compatible. If the resultant penalty functions are weighed to reflect properly their relative importance, then the criterion of plant performance will yield an optimum compromise when minimized (or maximized).

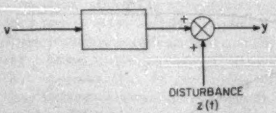
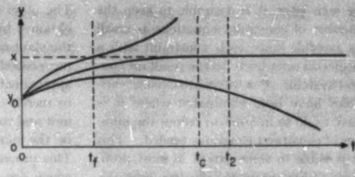


Fig. 4 Example of a system with a known disturbance, z(t)

Fig. 5. Selection of various points in a trial extrapolation



The reasons for introducing boundary conditions were stated and the relationship between the boundary condition at time t_1 and desired steady-state behavior was discussed. A set of specifications on steady-state behavior was established and then developed into a criterion of controller performance. It is clear that this criterion is reduced to zero when the desired steady-state conditions are reached. Once again, the form is arbitrary, reflecting only what is believed to be the desired steady-state condition. The most important feature of the criterion of controller performance is the fact that it is overspecified. In general, one can only specify the value of the various control variables and dependent variables. In a system of first-order equations such as describe the reactor, rates of change of variables usually may not be specified. The result is an overspecification of boundary conditions at t_1 and can be an incompatible set of conditions. If this situation exists, the controller will behave in such a fashion as to reduce the criterion of controller performance to a minimum. Consequently, careful weighting of the terms in the criterion of controller performance should result in a best possible compromise with regard to the steady-state conditions the control system finally selects.

The constraint equations were then discussed and a group of 11 constraint equations derived, the first seven directly from the original problem statement[15] and the remaining four by an approximation procedure. However, this approximation is probably acceptable if the rest of the processing units are behaving properly.

Next, the calculus of variations was applied and a partial solution resulted. If a complete solution were obtainable, the controller design would be finished. F_B, F_L, and H_w would be expressed as functions of plant conditions, desired steady-state conditions, and various constants previously specified. This would constitute a mathematical statement of the best way to manipulate these variables to minimize the criterion of plant operation and meet the steady-state requirements imposed.

Numerical Solution

The complete solution is analytically unattainable. Instead, a partial solution has resulted. Reference to Appendix I indicates that the solution is complete except for the inclusion of boundary conditions. In other words, the solution, equations 14–27, represents a sort of strategy or statement of how the control variables are to be changed as functions of time and plant conditions. It is only necessary to determine which initial control variable settings will result in the desired steady-state conditions at time t_1 if the strategy is followed.

As an example, assume a set of conditions prevails within the plant. Further, assume the control variables have each been set at some arbitrary value. The plant is then run according to its own dynamics, the control variables being manipulated according to equations 14–27. At time t_1 all the plant variables and control variables are measured. The resultant plant operation was optimum (I has been minimized) with respect to the boundary conditions measured at time t_1. The experiment will now be repeated. The same initial plant conditions are established, but different values for each control variable used. Again the plant is run in accordance with equations 14–27. At time t_1 all plant variables and control variables are measured. This set of measurements is found to be different from the set obtained in the first experiment. However, plant operation was again optimum with respect to this new set of boundary conditions measured at time t_1. This property of being optimum is guaranteed since a complete solution has been obtained between a given set of boundary values. The calculus of variations has guaranteed that either a maximum or minimum of I will be attained if a complete solution is obtained. However, a minimum must have been realized since a maximum may be obtained by making the control variables extremely large.

As can be seen, a unique set of initial control variable settings exists for each complete solution with respect to a given set of boundary conditions. It remains

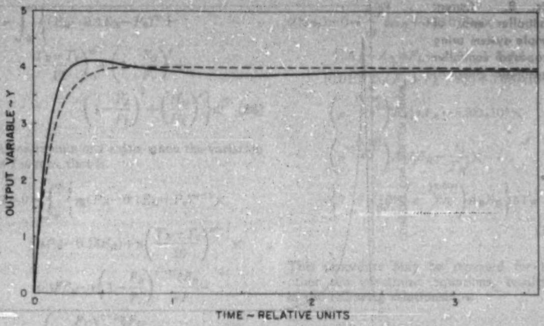


Fig. 6. Output of simple system using proposed controller as opposed to optimum output (quadratic criterion)

——— Response of proposed controller $n=2$
- - - - Optimum response

to find that set which relates to the complete solution with respect to the desired boundary conditions at time t_1, as given by the specification of controller performance. If it can be found, then a complete solution can be easily obtained with respect to the problem of interest, the control of the plant to the prescribed steady-state conditions.

Before discussing the procedures for obtaining a complete solution to the variational problem with respect to the prescribed boundary conditions at time t_1 by numerical means, it is useful to define a few terms.

"Path" will designate the time history of each plant variable associated with a given complete solution of the variational problem. For example, reactor temperature, T_R, may be said to describe a "path" (in time) between $t=0$ and $t=t_1$.

"Correct" will denote a characteristic of the complete solution to the variational problem associated with the prescribed boundary conditions at $t=t_1$. For example, "correct" paths may be said to exist, or "correct" initial variable settings have been selected.

Also, it will prove of more value in what follows to determine the correct set of initial values for the Lagrange multipliers, λ_i, used in equations 14–27 since they determine the control variables F_B, F_L, and H_w. In a similar manner, the derivative terms in the criterion of controller performance can be replaced by the sum of the squares of the rates of change of the Lagrange multipliers. Investigation of equations 14–27 shows that a steady-state condition is guaranteed if the Lagrange multipliers are not changing with time.

The experiments outlined previously describe an important method of finding the correct set of initial values for the Lagrange multipliers. A set of initial conditions for the Lagrange multipliers is surmised and, on a simulation of the plant, one determines how close to the desired boundary conditions one can come at $t=t_1$ for this first guess. Then the initial conditions are altered and a second experiment is made, etc. Hopefully, this process converges to the correct set of initial conditions for the Lagrange multipliers.

The computations proposed are, in fact, only trial extrapolations. Consider the extrapolation procedure. The extrapolation is being used to determine the unique set of initial values for the Lagrange multipliers which will cause the system to reach the desired boundary conditions at t_1. As stated before, this constitutes an optimum solution to the problem within the framework of the original problem. Therefore, t_1 must be specified. Inspection of equations 14–24 shows them to be dynamically unstable. This implies that some sets of initial values for the Lagrange multipliers may produce paths with convergent portions, but all paths must ultimately diverge from the desired steady-state conditions. What is sought is the path which touches the steady-state conditions before divergence commences. However, when the path will touch the desired final condition is unknown. Consequently, t_1 will be defined as the time of nearest approach to the desired final point. This is obtained by measuring Error of equation 2 continuously during the extrapolation and defining t_1 as the time when Error is a minimum. This is sufficient to indicate what particular set of initial values for the Lagrange multipliers is best with respect to the prescribed boundary condition at $t=t_1$. If Error ever is zero, then the guesses of these initial values constitute the unique, correct set. In general, this never happens.

The assumption that guessing will eventually lead to the correct set of initial values for the Lagrange multipliers implies that a technique is available for deriving new guesses for these values from existing information. In general, these techniques come under the general category of gradient procedures and methods of steepest descent. The efficiency or speed of convergence of these procedures usually depends on a knowledge of the surface being considered. Error measured at t_1 as a function of the initial values of the Lagrange multipliers describes some sort of hypersurface in 13-dimensional space. Unfortunately, very little is known about the nature of the hypersurfaces produced by the procedures that have been developed. As a consequence, rather inefficient descent procedures are currently being used to insure eventual convergence to a correct solution. This results in a phenomenally large number of extrapolations.

Once the computational procedure has obtained the correct complete solution to the variational problem, the control variables F_B, F_L, and H_w can be set according to the resultant initial conditions. However, it would be worthless to allow the system to operate with the control variables manipulated according to equations 14–27. This is an unstable policy and leads to trouble. Furthermore, computational errors and external upsets only make a bad situation worse. Instead the controller is required immediately to re-solve the problem starting from the last correct set of values. In this manner the controller is continuously updating its solution to the variational problem as plant conditions and disturbance conditions change. Controlling with unstable equations does seem undesirable. However, if it is realized that at a given instant of present time a path is sought which leads toward the desired final conditions, and that at the next instant of time such a path will again be sought, then, as long as any finite error exists, the system will always be approaching the desired steady-state conditions. Consequently, the controller that results from the unstable equations must, with the process, be stable if the controller can keep up with the controlled system in determining the correct set of initial values for the Lagrange multipliers. If the system is allowed to proceed along a given path for a finite time interval, it can conceivably proceed into the divergent portion of this path. However, the system is corrected the next time the problem is solved. The resultant behavior is a limit cycle. It is evident, then, that high-speed computation is important. As computational abilities improve, this use of the extrapolative technique will more closely approximate a continuous control system. Fig. 2 is a block diagram of the controller and its relationship to the plant.

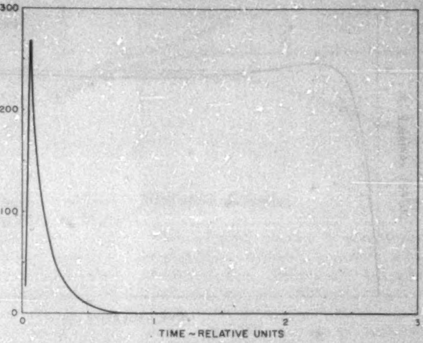


Fig. 8. Current controller error of simple system using proposed controller (quadratic criterion)

Simulation Results

Up to the present time, simulation efforts have been carried out in two areas only. Extremely simple systems have been controlled by the methods that have been proposed (see Appendix II) and the control systems for more advanced processing systems have been checked to assure that the computational procedure converges to the desired complete solution of the variational problem for selected cases. In all cases, the simulation is a digital computer program, the digital machine serving as controlled system and controller simultaneously. A complete program normally consists of a main program termed "Plant." This is as complete a simulation of the system to be controlled as can be accomplished. The main program is caused to exit to a subroutine called "Controller." This is the desired steepest descent procedure previously discussed. A second subroutine termed "Model" is available to "Controller." This is a second simulation of the plant and is used by Controller for the extrapolation computations which have been discussed above. Model is generally programmed so that one or more of a large variety of discrepancies can be caused to exist between Model and Plant at the time the programs are run in order to determine the effects of error in creating a mathematical model of the actual system which is to be controlled.

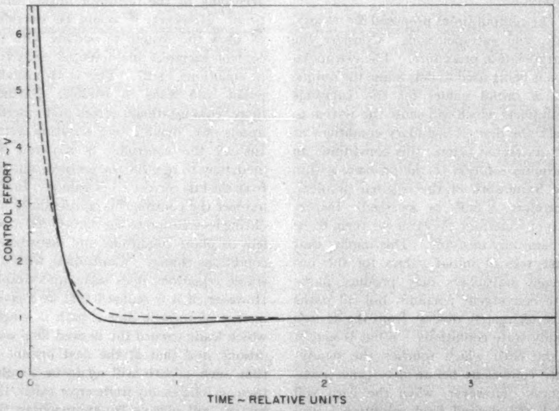


Fig. 7. Control effort of simple system using proposed controller as opposed to optimum control effort (quadratic criterion)

—— Control effort of proposed controller n = 2
- - - - Optimum control effort

These digital simulations have shown that:

1. The mode used for the extrapolation computations need be no more exact, as far as description of plant dynamics is concerned, than is currently attainable. Of course, such errors cause the actual path to deviate from the desired optimum, but stable, acceptable control can usually be accomplished.

2. The control system may require up to 200 or more extrapolations to achieve an acceptable solution when started from some arbitrary guesses for the initial values of the Lagrange multipliers in the case of a system as complex as that considered in this paper (11 Lagrange multipliers). However, if the procedure is started from guesses for the initial values of the Lagrange multipliers which are close to the correct initial values for these multipliers (as is the case during normal controller operation), the number of extrapolation computations can usually be reduced by a factor of about ten.

These conclusions are rather general and vary somewhat from case to case. Three cases have been considered by the authors[5] and consequently the data are too sparse to be more specific in terms of results of simulations of industrial processing units or complexes of units.

Appendix I. Partial Solution of Minimization of the Plant-Performance Criterion by Variational Calculus

This appendix treats the application of the calculus of variations to equation 1, the criterion of plant operation, and equations 3–13, the dynamic constraint equations, in order to obtain equations 14–27, the partial solutions of the problem of minimization of the criterion of plant performance (see references 1 and 2).

Equation 28 is the integral criterion to be minimized.

$$I=\int_0^{t_1}\Big\{(P_R-0.1E_R-P_0)^m+\frac{T_R-T_0}{10}\Big)^n+\left(1-\frac{F_B}{F_1}\right)^r+\left(1-\frac{F_L}{F_2}\right)^r+\left(\frac{H_w}{H_0}\right)^r\Big\}dt \quad (28)$$

An extremum of *I* exists when the variation of *I* is zero, that is,

$$\delta I=0=\int_0^{t_1}\Big\{m(P_R-0.1E_R-P_0)^{m-1}\times(\delta P_R-0.1\delta E_R)+n\left(\frac{T_R-T_0}{10}\right)^{n-1}\times 0.1\delta T_R-r\left(1-\frac{F_B}{F_1}\right)^{r-1}\frac{\delta F_B}{F_1}-r\left(1-\frac{F_L}{F_2}\right)^{r-1}\frac{\delta F_L}{F_2}+r\left(\frac{H_w}{H_0}\right)^{r-1}\frac{\delta H_w}{H_0}\Big\}dt \quad (29)$$

If all the variational terms in equation 29 were independent, it could be shown that all their coefficients must be identically zero for δI to be zero. However, this is not the case. Some variational quantities depend on others so that the solution of equation 29 is not quite so simple. This difficulty is readily overcome by using the constraint equations.

If equations relating the variables of equation 29 are available, they may be written:

$$\phi(\text{variables})=0 \quad (30)$$

This function can be multiplied by an arbitrary function of time without changing the equality. Since the function is zero, its variation must be zero and the integral of the variation must also be zero, that is,

$$\int_{10}^{T_1}\delta(\lambda\phi)dt=0 \quad (31)$$

Adding equation 31 to equation 29 does not alter equation 29 since a zero has merely been added. However, it allows the specification of λ such that the coefficient of a dependent variation is made to vanish. Inasmuch as λ is assumed to be completely arbitrary, this technique does not alter the validity of any equations. If a number of constraint equations are available, this process may be repeated for all the dependent variational terms that arise. When completed, only the coefficients of the independent variational terms remain. These may then be equated to zero separately because of the independence of these variational terms.

This technique shall now be applied. Rewriting equation 3, the first constraint equation, with the appropriate Lagrange multiplier, equation 32 is obtained.

$$\lambda_1\phi_1=0=\lambda_1\Big\{0.01078[F_A+F_LA_L-A_R(F_A+F_B+F_L)]-5.98\times 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)A_RB_R-\frac{dA_R}{dt}\Big\} \quad (32)$$

From this, the variation of $\lambda_1\phi_1$ may be obtained:

$$\delta(\lambda_1\phi_1)=0=\frac{d\lambda_1}{dt}\delta A_R+0.01078\lambda_1\times\{(A_L-A_R)\delta F_L-A_R\delta F_B+F_L\delta A_L\}-[0.01078\lambda_1(F_A+F_B+F_L)+5.98\lambda_1 10^9\times\left(e^{-\frac{12{,}000}{T_R}}\right)B_R]\delta A_R-5.98\lambda_1 10^9\times\left(e^{-\frac{12{,}000}{T_R}}\right)A_R\delta B_R-\frac{\lambda_1}{T_R^2}\times\Big\{7.17\times 10^{13}\left(e^{-\frac{12{,}000}{T_R}}\right)A_RB_R\Big\}\delta T_R \quad (33)$$

This procedure may be repeated for the other ten constraint equations, resulting in the following relationships:

$$\delta(\lambda_2\phi_2)=0=\frac{d\lambda_2}{dt}\delta B_R+0.01078\lambda_2\times\{(1-B_R)\delta F_B+(B_L-B_R)\delta F_L+F\delta_LB_L\}-\lambda_2[0.01078(F_A+F_B+F_L)+5.98\times 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)A_R+2.6\times 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)C_R]\delta B_R-5.98\lambda_2 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)B_R\delta A_R-2.6\lambda_2 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)B_R\delta C_R-\frac{\lambda_2}{T_R^2}\Big\{7.17\times 10^{13}\left(e^{-\frac{12{,}000}{T_R}}\right)A_RB_R+3.9\times 10^{16}\left(e^{-\frac{15{,}000}{T_R}}\right)B_RC_R\Big\}\delta T_R \quad (34)$$

$$\delta(\lambda_3\phi_3)=0=\frac{d\lambda_3}{dt}\delta C_R+0.01078\lambda_3\times\{F_L\delta C_L+(C_L-C_R)F\delta_L-C_R\delta F_B\}-\lambda_3[0.01078(F_A+F_B+F_L)+5.2\times 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)B_R+9.63\times 10^{15}\left(e^{-\frac{20{,}000}{T_R}}\right)P_R]\delta C_R+\lambda_3[11.96\times 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)A_R-5.2\times 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)C_R]\delta B_R+11.96\lambda_3 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)B_R\delta A_R-9.63\lambda_3 10^{15}\left(e^{-\frac{15{,}000}{T_R}}\right)C_R\delta P_R-\frac{\lambda_3}{T_R^2}\Big\{7.8\times 10^{16}\left(e^{-\frac{15{,}000}{T_R}}\right)B_RC_R+19.26\times 10^{19}\left(e^{-\frac{20{,}000}{T_R}}\right)C_RP_R-14.34\times 10^{13}\left(e^{-\frac{12{,}000}{T_R}}\right)A_RB_R\Big\}\delta T_R \quad (35)$$

$$\delta(\lambda_4\phi_4)=0=\frac{d\lambda_4}{dt}\delta E_R+0.01078\times\{(E_L-E_R)\delta F_L+F_L\delta E_L-E_R\delta F_B-(F_A+F_B+F_L)\delta E_R\}\lambda_4+5.2\lambda_4 10^{12}\left(e^{\frac{-15{,}000}{T_R}}\right)(B_R\delta C_R+C_R\delta B_R)+\frac{\lambda_4}{T_R^2}\Big\{7.8\times 10^{16}\left(e^{-\frac{15{,}000}{T_R}}\right)\times B_RC_R\Big\}\delta T_R \quad (36)$$

$$\delta(\lambda_5\phi_5)=0=\frac{d\lambda_5}{dt}\delta G_R-0.01078\lambda_5\times\{G_R(\delta F_B+\delta F_L)+(F_A+F_B+F_L)\delta G_R\}+14.45\lambda_5 10^{15}\left(e^{-\frac{20{,}000}{T_R}}\right)(C_R\delta P_R+P_R\delta C_R)+28.9\frac{\lambda_5}{T_R^2}10^{19}\times\left(e^{-\frac{20{,}000}{T_R}}\right)C_RP_R\delta T_R \quad (37)$$

$$\delta(\lambda_6\phi_6)=0=\frac{d\lambda_6}{dt}\delta P_R+0.01078\lambda_6\times\{0.1F_L\delta E_L+(0.1E_L-P_R)\delta F_L-P_R\delta F_B\}-\lambda_6[0.01078(F_A+F_B+F_L)+4.82\times 10^{15}\times\left(e^{-\frac{20{,}000}{T_R}}\right)C_R]\delta P_R+\lambda_6[2.6\times 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)B_R-4.82\times 10^{15}\left(e^{-\frac{20{,}000}{T_R}}\right)P_R]\delta C_R+2.6\lambda_6 10^{12}\left(e^{-\frac{15{,}000}{T_R}}\right)C_R\delta B_R+\frac{\lambda_6}{T_R^2}\Big\{3.9\times 10^{16}\left(e^{-\frac{15{,}000}{T_R}}\right)B_RC_R-9.64\times 10^{19}\left(e^{-\frac{20{,}000}{T_R}}\right)C_RP_R\Big\}\delta T_R \quad (38)$$

$$\delta(\lambda_7\phi_7)=0=\frac{d\lambda_7}{dt}\delta T_R+\lambda_7\Big\{[4{,}480\times 10^{13}\left(e^{-\frac{12{,}000}{T_R}}\right)A_RB_R+1{,}460\times 10^{16}\left(e^{-\frac{15{,}000}{T_R}}\right)B_RC_R+10{,}320\times 10^{19}\left(e^{-\frac{20{,}000}{T_R}}\right)C_RP_R]\frac{1}{T_R^2}-0.01078(F_A+F_B+F_L)\Big\}\delta T_R-0.01078\lambda_7\{(T_R-T_L)\delta F_L+(T_R-T_B)\}+3{,}730\lambda_7 10^9\times\left(e^{-\frac{12{,}000}{T_R}}\right)B_R\delta A_R+5{,}160\lambda_7 10^{15}\times\left(e^{-\frac{20{,}000}{T_R}}\right)C_R\delta P_R+\lambda_7[3730\times 10^9\left(e^{-\frac{12{,}000}{T_R}}\right)A_R+975\times 10^{12}\times$$

$$\left(e^{-\frac{15,000}{T_R}}\right)C_R\delta B_R+\lambda_7\delta H_w+$$

$$\lambda_7[975\times 10^{13}\left(e^{-\frac{15,000}{T_R}}\right)B_R+5,160\times$$

$$10^{15}\left(e^{-\frac{20,000}{T_R}}\right)P_R]\delta C_R \quad (39)$$

$$\delta(\lambda_8\phi_8)=0=\frac{d\lambda_8}{dt}\delta A_L+2.665\lambda_8\delta A_R-2.425\lambda_8\delta A_L \quad (40)$$

$$\delta(\lambda_9\phi_9)=0=\frac{d\lambda_9}{dt}\delta B_L+2.665\lambda_9\delta B_R-2.425\lambda_9\delta B_L \quad (41)$$

$$\delta(\lambda_{10}\phi_{10})=0=\frac{d\lambda_{10}}{dt}\delta C_L+2.665\lambda_{10}\delta C_R-2.425\lambda_{10}\delta C_L \quad (42)$$

$$\delta(\lambda_{11}\phi_{11})=0=\frac{d\lambda_{11}}{dt}\delta E_L+2.665\lambda_{11}\delta E_R-2.425\lambda_{11}\delta E_L \quad (43)$$

These zeros are added to the expression for the variation of I.

$$\delta I=\int_0^{t_1}\Bigg\{m(P_R-0.1E_R-P_0)^{m-1}\times(\delta P_R-0.1\delta E_R)+n\left(\frac{T_R-T_0}{10}\right)^{n-1}\times 0.1\delta T_R-r\left(1-\frac{F_B}{F_1}\right)^{r-1}\frac{\delta F_B}{F_1}-r\left(1-\frac{F_L}{F_2}\right)^{r-1}\frac{\delta F_L}{F_2}+r\left(\frac{H_w}{H_0}\right)^{r-1}\times\frac{\delta H_w}{H_0}+\sum_{i=1}^{11}\delta(\lambda_i\phi_i)\Bigg\}dt \quad (44)$$

Multipliers are selected such that the coefficients of the variations of the 11 dependent variables are zero. This is allowable since the Lagrange multipliers are arbitrary. Summing all the coefficients of δA_R and solving for $d\lambda_1/dt$, one obtains:

$$\frac{d\lambda_1}{dt}=5.98\times 10^9\left(e^{-\frac{12,000}{T_R}}\right)B_R(\lambda_1+\lambda_8-2\lambda_3-625\lambda_7)+0.01078\lambda_1F_R-2.665\lambda_8 \quad (45)$$

This will be recognized as equation 14. Summing all coefficients of δB_R and solving for $d\lambda_2/dt$ yields equation 15. Equations 16–24 can be obtained similarly. The only terms in equation 49 that have not been considered are those involving variations of control variables. These variations are entirely independent of each other so that the coefficient of each of these quantities must be zero independently of the other remaining terms. Summing the coefficients on δH_w and equating to zero yields equation 25. In a similar manner, equations 26 and 27 are obtained by consideration of the coefficients of δF_L and δF_B respectively.

If these equations are all satisfied, then the resultant solution represents an optimum path with respect to equation 1. Methods discussed in this paper insure that this optimum path satisfies the end-condition requirements.

Appendix II. Control of Simple Systems

This appendix treats the control of simple systems. The Euler equation is derived in its standard form. Following this are examples of the control of two first-order systems; the first is nonlinear, showing the general applicability of the technique, while the second example is a linear system with a quadratic performance criterion. In the latter case analytical solutions are obtainable which are compared with solutions obtained by simulating an extrapolating controller of the type proposed in this paper. Finally, a nonquadratic performance criterion is considered and the transients obtained from the simulation of the system and its controller are discussed.

Derivation of the Euler Equation

Inasmuch as the proposed technique is restricted to integral criteria of system performance, consider such a criterion,

$$I=\int_{t_1}^{t_2}R\,dt \quad (46)$$

with the requirement that I be minimized over the time interval t_1 to t_2. R is an arbitrary function of quantities in the controlled system. It is also possible to include quantities that are normally considered as external to the system. The boundaries of the system under consideration can then simply be extended to include such quantities.

If the system under consideration has n variables, $x_1, x_2, \ldots x_n$, which are functions of time, then R is a function of these variables and their derivatives.

$$R=R[t, x_1, x_2, \ldots x_n; \dot{x}_1, \dot{x}_2, \ldots \dot{x}_n; \ddot{x}_1, \ldots \ddot{x}_n; \text{etc.}] \quad (47)$$

These variables may or may not depend upon each other. Suppose that unique functions of time exist for each variable that minimizes I. These functions are denoted by $x_{10}(t)$, $x_{20}(t)$, etc. These functions satisfy the boundary conditions at t_1 and t_2. A small variation in each of these functions is introduced.

$$x_1(t)=x_{1_0}(t)+\epsilon\eta_1(t)$$
$$x_2(t)=x_{2_0}(t)+\epsilon\eta_2(t) \quad (48)$$

etc , where

$$\sigma_1(t_1)=\eta_1(t_2)=\eta_2(t_2)=0$$

Likewise

$$\dot{x}_1(t)=\dot{x}_{10}(t)+\epsilon\dot{\eta}_1(t)$$
$$\dot{x}_2(t)=\dot{x}_{20}(t)+\epsilon\dot{\eta}_2(t) \quad (49)$$

Equation 49 is substituted into the expression for I, giving

$$I(\epsilon)=\int_{t_1}^{t_2}R[t, x_{10}+\epsilon\eta_1, x_{2_0}+\epsilon\eta_2, \ldots; \dot{x}_{1_0}+\epsilon\dot{\eta}_1, \dot{x}_{2_0}+\epsilon\dot{\eta}_2\ldots]dt \quad (50)$$

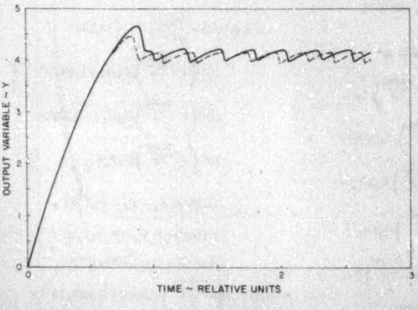


Fig. 9. Output of simple system using proposed controller for two conditions of controller computing speed (nonquadratic criterion)

——System response for n=10

- - - -System response for n=10 and doubled controller computing speed

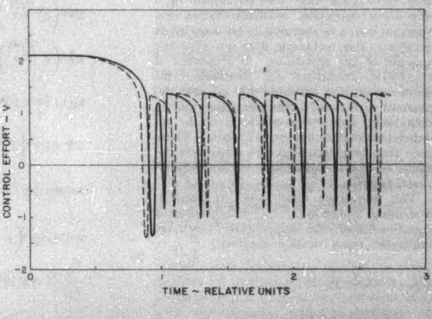


Fig. 10. Control effort of simple system using proposed controller for two conditions of controller computing speed (nonquadratic criterion)

——Control effort for n=10

- - - -Control effort for n=10 and doubled controller computing speed

If I is to be a minimum it is necessary, but not sufficient, that

$$\left.\frac{dI(\epsilon)}{d\epsilon}\right|_{\epsilon=0}=0 \qquad (51)$$

Since ϵ is independent of time,

$$\frac{dI}{d\epsilon}=\int_{t_1}^{t_2}\frac{dR}{d\epsilon}\,dt \qquad (52)$$

$$\frac{dR}{d\epsilon}=\frac{\partial R}{\partial x_1}\eta_1+\frac{\partial R}{\partial x_2}\eta_2+\ldots+\frac{\partial R}{\partial \dot{x}_1}\dot{\eta}_1+\frac{\partial R}{\partial \dot{x}_2}\dot{\eta}_2+\ldots+\frac{\partial R}{\partial \ddot{x}_1}\ddot{\eta}_1+\ldots$$

$$=\sum_{i=1}^{n}\frac{\partial R}{\partial x_i}\eta_i+\sum_{n=1}^{n}\frac{\partial R}{\partial \dot{x}_i}\dot{\eta}_i+\sum_{i=1}^{n}\frac{\partial R}{\partial \ddot{x}_i}\ddot{\eta}_i+\ldots \qquad (53)$$

$$\int_{t_1}^{t_2}\frac{dR}{d\epsilon}\,dt=\sum_{i=1}^{n}\left\{\int_{t_1}^{t_2}\frac{\partial R}{\partial x_i}\eta_i dt+\int_{t_1}^{t_2}\frac{\partial R}{\partial \dot{x}_i}\dot{\eta}_i dt+\int_{t_1}^{t_2}\frac{\partial R}{\partial \ddot{x}_i}\ddot{\eta}_i dt+\ldots\right\} \qquad (54)$$

Integrating

$$\int_{t_1}^{t_2}\frac{\partial R}{\partial \dot{x}_i}\dot{\eta}_i dt$$

by parts, one obtains:

$$\int_{t_1}^{t_2}\frac{\partial R}{\partial \dot{x}_i}\dot{\eta}_i dt=\left.\frac{\partial R}{\partial \dot{x}_i}\eta_i\right|_{t_1}^{t_2}-\int_{t_1}^{t_2}\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{x}_i}\right)\eta_i dt \qquad (55)$$

But

$$\eta_i(t_1)=\eta_i(t_2)=0$$

so

$$\left.\frac{\partial R}{\partial \dot{x}_i}\eta_i\right|_{t_1}^{t_2}=0 \text{ for all } i \qquad (56)$$

Integrating

$$\int_{t_1}^{t_2}\frac{\partial R}{\partial \ddot{x}^i}\ddot{\eta}_i dt$$

twice by parts:

$$\int_{t_1}^{t_2}\frac{\partial R}{\partial \ddot{x}_i}\ddot{\eta}_i dt=\left.\frac{\partial R}{\partial \ddot{x}_i}\dot{\eta}_i\right|_{t_1}^{t_2}-\int_{t_1}^{t_2}\frac{d}{dt}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)\dot{\eta}_i dt=\left.\frac{\partial R}{\partial \ddot{x}_i}\dot{\eta}_i\right|_{t_1}^{t_2}-\left.\frac{d}{dt}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)\eta_i\right|_{t_1}^{t_2}+\int_{t_1}^{t_2}\frac{d^2}{dt^2}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)\eta_i dt \qquad (57)$$

With the requirement that η_i and $\dot{\eta}_i$ be zero at t_1 and t_2:

$$\int_{t_1}^{t_2}\frac{\partial R}{\partial \ddot{x}_i}\ddot{\eta}_i dt=\int_{t_1}^{t_2}\frac{d^2}{dt^2}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)\eta_i dt \qquad (58)$$

This process continues for the integration of higher order terms in a similar manner. Substituting into equation 54 there results

$$\int_{t_1}^{t_2}\frac{dR}{d\epsilon}\,dt=0=\sum_{i=1}^{n}\int_{t_1}^{t_2}\left[\frac{\partial R}{\partial x_i}-\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{x}_i}\right)+\frac{d^2}{dt^2}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)-\ldots\right]\eta_i dt \qquad (59)$$

Since η_i is arbitrary, the quantity in the brackets must be zero.[1] Using a more common notation, η_i is the variation of x_i, or

$$\eta_i=\delta x_i \qquad (60)$$

In general, the variations of the variables (x_i) are not independent, that is, x_j is a function of x_k so that δx_j is not independent of δx_k. In this situation the summation sign must be brought back inside the integration of equation 59 giving

$$\sum_{i=1}^{n}\left[\frac{\partial R}{\partial x_i}-\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{x}_i}\right)+\frac{d^2}{dt^2}\left(\frac{\partial R}{\partial \ddot{x}_i}\right)-\ldots\right]\delta x_i=0 \qquad (61)$$

where δx_i must be expressed as some function of δx_k.

Equation 61, the Euler equation, is a necessary but not sufficient condition that I be a minimum. Therefore, if the control system is an optimum system, equation 61 must be satisfied.

A Simple Nonlinear System

In order to try this approach, consider a very simple nonlinear example. A positioning device has an output y and input v. The dynamic behavior is given by

$$v=y+(\dot{y})^3 \qquad (62)$$

Furthermore, $|v|$ cannot exceed a limiting value v_{max}.

$$|v|\leq v_m \qquad (63)$$

or

$$-v_m\leq v\leq v_m \qquad (64)$$

It is required that y be held at a specified value x in the presence of a disturbance that is not specified. Therefore, this requires that y return to x after any disturbance in a manner which minimizes an integral function of the error, $x-y$. For the purpose of this example the integral criterion is

$$\int(x-y)^4dt$$

Inasmuch as v has certain limitations, it is necessary to include this effect in the criterion of system performance in such a manner that the criterion becomes enormously large if v attempts to exceed v_m but is unaffected otherwise. Thus the performance criterion is modified to be

$$I=\int_0^{t_2}\left[(x-y)^4+\left(\frac{v}{v_m}\right)^n\right]dt \qquad (65)$$

where n is a large even integer in order that

$$\left(\frac{v}{v_m}\right)^n\geq 0$$

for any v. If n is sufficiently large, then the second term of the criterion is small for $|v|<v_m$. As $n\rightarrow\infty$, the term becomes discontinuous at $v=v_m$. However, because of the very large increase in I when $|v|$ exceeds v_m for a short period of time, it is reasonable to assume that the time integral of such excesses will tend to be minimized by the optimization procedure. The time integral of v is the quantity that changes the value of y in the system of equation 62. Consequently, a simple limiter will ensure that v_m is not exceeded for extremely short time intervals and the analysis is approximately correct since it is known that the criterion is severely penalized if $|v|>v_m$ for any finite time interval. Therefore,

$$R=(x-y)^4+\left(\frac{v}{v_m}\right)^n \qquad (66)$$

Using equation 62,

$$R=(x-y)^4+\left(\frac{y+\dot{y}^3}{v_m}\right)^n \qquad (67)$$

Since x is a constant,

$$\delta x=0 \qquad (68)$$

Furthermore, no control over time can be exercised,

$$\delta t=0 \qquad (69)$$

Equation 69 is true in general and is not a consequence of this particular problem. Also,

$$R=R[x,\,t,\,y,\,\dot{y}] \qquad (70)$$

Applying equation 61,

$$\left[\frac{\partial R}{\partial y}-\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{y}}\right)\right]\delta y=0$$

or

$$\frac{\partial R}{\partial y}-\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{y}}\right)=0 \qquad (71)$$

Performing the indicated differentiations, and using equation 67,

$$\frac{\partial R}{\partial y}=-4(x-y)^3+n\left(\frac{y+\dot{y}^3}{v_m}\right)^{n-1}\left(\frac{1}{v_m}\right)=-4(x-y)^3+\frac{n}{v_m}\left(\frac{v}{v_m}\right)^{n-1} \qquad (72)$$

$$\frac{\partial R}{\partial \dot{y}}=n\left(\frac{y+\dot{y}^3}{v_m}\right)^{n-1}\left(\frac{3}{v_m}\right)\dot{y}^2 \qquad (73)$$

$$\frac{d}{dt}\left(\frac{\partial R}{\partial \dot{y}}\right)=\frac{3n}{v_m}\dot{y}^2(n-1)\left(\frac{y+\dot{y}^3}{v_m}\right)^{n-2}\times\left(\frac{1}{v_m}\right)\frac{d}{dt}(y+\dot{y}^3)+n\left(\frac{v}{v_m}\right)^{n-1}\frac{3}{v_m}2\dot{y}\ddot{y}$$

$$=\frac{3n}{v_m}\left\{\left(\frac{n-1}{v_m}\right)\dot{y}^2\left(\frac{v}{v_m}\right)^{n-2}\dot{v}+\left(\frac{v}{v_m}\right)^{n-1}2\dot{y}\ddot{y}\right\} \qquad (74)$$

Substituting into equation 71 and solving for $\dot{v}$:

$$-4(x-y)^3+\frac{n}{v_m}\left(\frac{v}{v_m}\right)^{n-1}=\frac{3n}{v_m}\left(\frac{v}{v_m}\right)^{n-2}\times\dot{y}\left\{\left(\frac{n-1}{v_m}\right)\dot{y}\dot{v}+2\ddot{y}\right\} \qquad (75)$$

$$\dot{v}=\frac{v_m}{(n-1)\dot{y}}\left\{-2\ddot{y}+\left(\frac{v_m}{v}\right)^{n-2}\frac{v_m}{3n\dot{y}}\times\right.$$
$$\left.\left[-4(x-y)^3+\frac{n}{v_m}\left(\frac{v}{v_m}\right)^{n-1}\right]\right\} \qquad (76)$$

Equation 76 is a law for altering $\dot{v}$ as a function of v, y, and its derivatives. Examination shows, however, that to attempt to control the system by such a law would lead to violent instability. It is possible that equation 76 represents a control law to maximize the criterion I rather than to minimize it. A little thought indicates that a worse condition could exist by setting $v=v_m$ and leaving the system alone. This means that equation 76 is not a maximizing control law. A paradox has apparently risen. Clearly, equation 76 is not a control law to maximize I. It is also equally clear that it does not minimize I since instability exists. Since it was derived from equation 71, it must either maximize, minimize, or represent a saddle. One must examine the characteristics of equation 76 more carefully.

Assume that $y=x-1$ at $t=0$. What path will y take as time increases? It is apparent that the path taken depends on the initial conditions on v. Approximate forms of these paths for various values of v_0 (the initial condition on v) are shown in in Fig. 3. One of the paths for a particular value of v_0, denoted by v_1, gives a time variation of y that ultimately reaches $y=x$ and stays there. This situation is an unstable equilibrium. However, equation 76 describes the desired variation of v with time to minimize I if $v=v_1$ is the initial condition. Equation 76 therefore, represents a saddle point and the correct value of v_1 must be found to insure that the correct path from the saddle point is taken. In general, equation 76 can yield nonoptimum solutions unless the correct value of v_1 is chosen.

The next question that arises is: If this is the desired path, how can one be sure that the system will follow this path in the presence of further disturbance? The answer is that the system will not do so, and actually one would not wish it to. The design was originally posed as the problem of recovering from a disturbance. It was implicitly assumed that no further disturbances would occur when the disturbance was defined as unspecified. If it is specified, it must be added as an additional input. This is illustrated in Fig. 4. Including the disturbance will give a different control law. Also, once the system has deviated from a given path, it would seem much more reasonable to find a new optimum path related to existing conditions, than to attempt to recover an obsolete solution. Certainly, if a new disturbance occurs, a new set of initial conditions is set up and a new optimum path must be found. This path is determined if the old control law is applied (equation 76) with a new initial condition on v to give the desired optimum path. Now the problem becomes that of determining the correct initial condition on v.

In the original derivation of equation 59, it was assumed that conditions at t_2 were known. Since then, conditions at t_2 have been more or less ignored. In fact, a time t_2 must be defined, and the conditions to be satisfied at t_2 must be specified in the light of the control objectives. The specification of conditions at t_2 is the only means of determining a unique path and therefore a unique initial condition on v.

The specification of some time t_2 and the conditions which shall exist at that time is rather arbitrary as long as the specification uniquely determines the correct initial condition on v. Any time, t_2, and any specification of values at that time will satisfy the requirements of the original derivation of equation 66. What is desired is to cause y to move to a value x under the conditions stated. The simplest way to specify t_2 and the conditions at t_2 is to set

$$t_2=\infty$$

$$y=x$$

$$\dot{y}=\ddot{y}=\ldots=0$$

This certainly gives a correct solution, but it needlessly complicates the problem. Referring to Fig. 5, the problem can be considered essentially solved at some time represented by t_c. If the path is correct, then $y=x$ at t_c and the derivatives of y are approximately zero. However, the solution time is generally a function of all the conditions at $t=0$ and consequently is relatively difficult to determine. To resolve the difficulty, t_2 can be chosen well beyond t_c so as to guarantee that specifying $y=x$ gives the correct value of the initial condition on v. If conditions are specified at an earlier time such as t_f, the determination of these conditions is fairly involved. Picking t_2 as shown in Fig. 5 seems to be the simplest procedure. This method is an alternative to the method discussed in the body of the paper.

If, however, this is indeed a final-value controller and it is important that $y=x$ at t_f, then the simple specification that $y=x$ at $t=t_f$, is sufficient to find the v_0 corresponding to the path passing through x at t_f if such a path is realizable. The path will also minimize I over the time interval 0 to t_f commensurate with the stated boundary conditions. As can be seen, selection of t_2 is up to the designer and depends on the problem at hand.

Choosing t_2 as shown in Fig. 5, how does one go about finding the corresponding v_0? In general, v_0 cannot be found in closed analytical form. One approach is to integrate equation 76 and solve for the constant of integration so that $y=x$ at $t=t_2$. Then solve for v as a function of x, y, and all pertinent derivatives. If it were possible to do this, it would be found that v is a stable function of system conditions. This integration can be performed in extremely simple linear cases, in which cases v is stable. Furthermore, specification of t_2 a short time beyond t_c insures that derivatives of y and v can be zero at time t_2. A study of Fig. 5 will indicate that the path is flat from t_c to t_2 so that all derivatives must be approximately zero. Instability may cause computational trouble in this range however.

A Linear System

It is possible to obtain analytic solutions to the problem of the design of dynamic control systems when linear systems are considered and quadratic criteria of system performance are used.[3] Consequently, it is of interest to consider the use of the technique proposed in this paper to control such a system for the purpose of comparing the behavior of the resultant controller with control systems developed by more common techniques.

The system considered is

$$v=a\frac{dy}{dt}+by \qquad (77)$$

The criterion of system operation is

$$I=\int_0^{t_1}\left\{(y-3)^2+\left(\frac{v}{v_0}\right)^n\right\}dt \qquad (78)$$

Applying the technique of Appendix I, the following results are obtained:

$$v=v_0\left(\frac{\lambda v_0}{n}\right)^{\frac{1}{n-1}} \qquad (79)$$

$$\frac{d\lambda}{dt}=\frac{b}{a}\lambda+\frac{2}{a}(y-z)$$

This is the control law. If $n=2$, equation 79 may be easily integrated.[4] The criterion of controller performance is

$$\text{Error}=(y-z)^2+\left(\frac{dy}{dt}\right)^2+\left(\frac{d\lambda}{dt}\right)^2 \qquad (80)$$

From the derivation of the Euler equation, it is evident that only y may be specified at $t=t_1$. As a consequence, equation 80 is an overspecification of boundary conditions (see body of paper). However, a condition

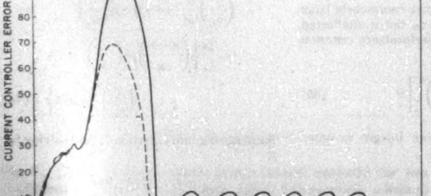


Fig. 11. Current controller error of simple system using proposed controller (nonquadratic criterion)

——— Current controller extrapolation error for n=10

– – – – Current controller extrapolation error for n=10 and doubled controller computing speed which causes Error to be zero is indeed attainable since

$$v = bz \qquad (81)$$

will eventually cause Error to approach zero. The problem must have a solution in the form of an optimal trajectory in time.

The system was assumed to have the following parameters for the purpose of simulation:

$a = 0.25$
$b = 0.25$
$v_0 = 2.0$
$z = 4.0$
$y = 0$ at $t = 0$
$n = 2$

The analytic solution to this set of conditions (the step response) is

$$y = 4(1 - e^{-t/r})$$

$$v = 1 + \frac{1-r}{r} e^{-t/r} \qquad (82)$$

where r is a form of Lagrange multiplier used to limit v.[4] The system of equation 77 and the controller designed by means of the techniques proposed in this paper was simulated digitally. The digital simulation used time increments of 0.025 time unit. The controller was allowed only six extrapolations per system step (of 0.025 time unit). The resultant time history of y is plotted in Fig. 6. Equation 82 is included for comparison. Fig. 7 compares the control effort of the extrapolating controller with that of equation 82. Fig. 8 is the value of Error currently being achieved by the control system. Observe that six extrapolations are nowhere near sufficient to solve the problem. However, even in the face of these large computing errors, the system is controlled acceptably well.

Two additional step responses were obtained from the simulation. In both these cases, $n = 10$. The previous responses have shown that the use of a quadratic criterion is somewhat ineffectual in limiting control effort excursions. Using $n = 10$ is an effort to determine the effectiveness of a stiffer penalty function. The step responses were taken with varying numbers of extrapolations available to the controller. In the first case, six extrapolations per system step were available. In the second case, this was raised to 11 extrapolations per system step. Figs. 9–11 show the responses in y, v, and Error, respectively, for the two cases. Clearly, a limit cycle has resulted from the inability of the controller to handle the computations. Also, this limit cycle is due to some extent to the discrete approximation used. This is emphasized by the fact that control effort, v, becomes strongly negative for only one time increment. It seems reasonable that the limit cycle would be reduced if the time increment were reduced. At present, this hypothesis has not been tested. The extremely large computational errors caused by the severe limitations placed on controller computations has not, however, caused complete breakdown of system behavior. The system runs and is forced to approach the desired condition. It is reasonable to assume that better responses would be obtained as the number of extrapolations approaches the continuous control situation obtained.

Bibliography

Calculus of Variations

1. Methods of Applied Mathematics (book), F. B. Hildebrand. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1952, chap. 2.

2. Calculus of Variations (book), R. Weinstock. McGraw-Hill Book Company, Inc., New York, N. Y., 1952.

Use of Calculus of Variations in Control

3. Optimum Non-linear Controllers, W. Kipiniak. *Report 7793-R-2*, Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1958, chaps. 3–5.

4. Analytical Design of Linear Feedback Controls (book), G. C. Newton, L. A. Gould, V. F. Kaiser. John Wiley & Sons, Inc., New York, N. Y., 1957.

5. Synthesis of an Optimum Distillation Controller, A. E. Beecher. *E.E. Thesis*, Massachusetts Institute of Technology, June 1960.

Miscellaneous

6. Dynamic Methods of Bubble Plate Analysis, L. M. Schwartz. *Report 7793-R-4*, Electronic Systems Laboratory, Massachusetts Institute of Technology, 1959.

7. Control System Synthesis (book), J. C. Truxal. McGraw-Hill Book Company, Inc., 1955, chap. 11.

8. Stability of Optimum Nonlinear Controllers, J. Rodriguez. *B.S. Thesis*, Massachusetts Institute of Technology, June 1960.

9. An Analysis of Transients in Mass-Transfer Operations, H. M. Teager. *Sc.D. Thesis*, Massachusetts Institute of Technology, 1955.

10. A Generalized Chemical Processing Model for the Investigation of Computer Control, Theodore J. Williams, Robert E. Otto. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, Nov. 1960, pp. 458–73.

Design of High Coercivity Permanent Magnets, Exposed to External Fields

M. J. SCHINDLER
NONMEMBER AIEE

Synopsis: This paper describes a method for designing permanent magnets that are to be exposed to magnetic fields from external sources. Under such conditions, a widely used procedure for magnet design cannot be relied upon to give accurate results. It it shown that this procedure does not conform generally to the basic equations of magnetism and, therefore, can be used only under limited circumstances. Engineers, not specialized in magnetics, are generally unware of these limitations.

Because the demagnetizing field is a function of intensity of magnetization (intrinsic induction) rather than induction, the demagnetization factor ("load line") operates in an **I-H** plane while for the demagnetization curve the **B-H** plane is generally used. These ordinate systems may be interchanged if the contribution of the magnetic field to the induction is negligible. Modern (ceramic) materials, however, exhibit coercive forces in the same order of magnitude as the saturation induction. Consequently, the improper use of the "external" magnetization curve (**B-H** curve), yields demagnetizing fields considerably lower than those found experimentally. The use of the "intrinsic" demagnetization curve (**I-H** curve), as outlined in this paper, leads to results which are in excellent agreement with experiments.

A MAGNETIC system was recently designed in which several magnets of oriented barium ferrite were stacked together to form a magnet assembly. During assembly these magnets became partially demagnetized, although the usual graphical design approach indicated that they should work well within the straight part of the demagnetization curve. The theoretical and experimental investigations described below were initiated as a result of this observation.

When the graphical design procedure mentioned previously is critically checked against basic requirements of magnetic theory, it becomes immediately obvious that the **B-H** curve is improperly used. If permanent-magnet problems are treated magnetostatically, instead of in terms of an electrodynamic analogy, no such errors will be possible, and better accuracy may be expected. This view has been confirmed by experimental results and is shown below on a design example.

Conventional Design Procedure

The design procedure for permanent magnets is so well established that all

---

Paper 61-194, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 5, 1959; made available for printing December 8, 1960.

M. J. Schindler is with the Radio Corporation of America, Harrison, N. J.

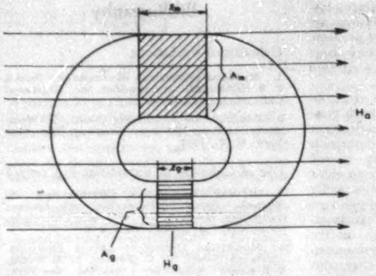


Fig. 1. Simple magnetic circuit exposed to an external field

textbooks on the subject are almost identical. For the simplest case with no stray flux conventional design proceeds as follows:

1. The magnetomotive force M in the gap equals that in the magnet:

$$M=H_g l_g=H_m l_m \tag{1}$$

where H_g is the gap field, l_g is the gap length, H_m is the field in the magnet, and l_m is the length of the magnet, as shown in Fig. 1.

2. The flux in the gap equals the flux in the magnet

$$\phi=A_g H_g=A_m B_m \tag{2}$$

where A_g and A_m are the cross-sectional areas of the gap and the magnet, respectively, and B_m is the magnetic flux density in the magnet.

The division of equation 2 by equation 1 gives the equation for the load line λ as follows:

$$-\lambda=\frac{B_m}{H_m}=\frac{A_g}{A_m}\frac{l_m}{l_g} \tag{3}$$

Fig. 2 shows the application of the load line to the demagnetization curve of an oriented barium-ferrite material, and the location of a working point P. So far the procedure is in agreement with the theory.

According to general practice[1,2,3] the presence of an applied field, such as H_a in Fig. 1, displaces the load line by the amount H_a on the H axis, as shown in Fig. 2. In the example shown, this displacement would move the point P to a position just above the knee of the curve and would cause no loss in the efficiency of the magnet. This step, probably because of its simplicity, is generally not analyzed mathematically. However, it proves to be incorrect, as shown subsequently.

DEMAGNETIZATION FACTOR

For a given geometrical configuration, the load line ascribes a certain demagnetized field to a certain value of induction. Because the load line is a function of circuit geometry, it may naturally be considered as only another expression for the demagnetization factor, which occupies an important place in the theory of magnetism. In a general case, as shown in Fig. 3, the field in the magnet, H_m, is given by

$$H_m=H_a-H_d=H_a-NI \tag{4}$$

where H_a is the applied field, H_d is the demagnetizing field, N is the demagnetization factor, and I is the intensity of magnetization, defined by $4\pi I=B-H$. For the purpose of this treatment, it is simpler to use the quantity $4\pi I$ as a new unit I', which is often referred to as intrinsic induction and is given by

$$I'=B-H \tag{5}$$

Equation 4 shows that the demagnetizing field is a function of the intrinsic induction I', and not of the external induction B. It is now possible to define a "modified load line" λ^* for zero applied field ($H_m=H_d$) as follows:

$$-\lambda^*=\frac{I'}{H_m}=\frac{B_m-H_m}{H_m}=-\lambda-1, \text{ or } \lambda^*=\lambda+1 \tag{6}$$

This load line must be applied to the intrinsic demagnetization curve, which can be obtained from the initial curve by adding $-H$ to B. This procedure, shown in Fig. 4, yields the same demagnetizing field obtained with the standard procedure. This result of course is necessary inasmuch as equations 1 and 2 are correct.

However, the agreement between the two procedures described in this and the previous paragraphs is lost when an external field is applied. If, again, the applied field H_a in equation 4 is added by shifting the load line in Fig. 4, the load line now intersects well below the knee of the demagnetization curve. After the external field is removed, the working point now shifts from P to P', thus indicating a loss in the effectiveness of the magnet.

DANGER OF ANALOGIES

It seems that the deviation of the design method for permanent magnets from the lawful path was caused in part by the desire to press magnetic phenomena into electrical analogies. Faraday's concept of closed "lines of induction" was very helpful in making some properties of magnetism understandable, by likening them to flow.[4] In the course of time, however, Faraday's "concept" was transformed into physical "fact," and students to this day are taught Ohm's law of magnetism as the basis of magnetic engineering. Although this electrodynamic analogy may help a student understand magnetic circuits, which contain only nonretentive materials, permanent magnets do not lend themselves to such treatment without adverse side effects.

First, it must be understood that in a permanent magnet the field does not cause the flux, as a voltage causes a current. The field is just a by-product of the magnetization. If a cause has to be assigned, the molecular magnetization, practically

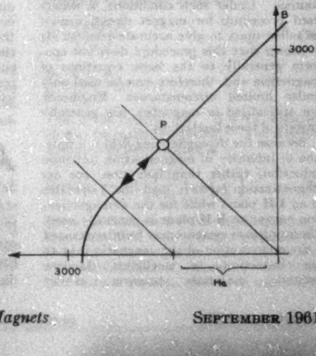


Fig. 2. Conventional design construction, using the B-H curve

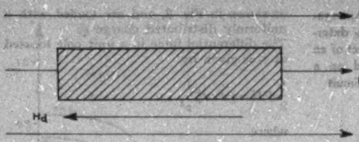


Fig. 3 (left). The demagnetizing field as used in equation 4

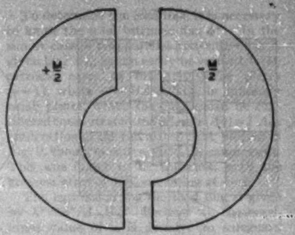


Fig. 5. Static equivalent of the "circuit" of Fig. 1

expressed in terms of intrinsic induction, must be considered. Furthermore, nothing "flows" in a magnetic circuit to justify an electrodynamic analogy. If an electrical analogy is needed, the electrostatic case should be considered, to which permanent magnetism is more closely related than to an electric circuit. This will also permit use of the physically meaningful magnetization I' instead of the induction B.

MAGNETOSTATIC DESIGN PROCEDURE

In the light of this reasoning, the "circuit" of Fig. 1 can be transformed into the simple geometric configuration of Fig. 5. If it is assumed that the permeability of the pole pieces is very high compared with that of the environment (which is true in almost any instance), the problem reduces to the determination of the field configuration between two equipotential surfaces, like charged conductors. Although there is a considerable "stray field" between the conductors, this field will be neglected for the moment to keep the comparison with the previously used "circuit" approach valid. The two gaps then show fields having the following values:

$$H_m = \frac{M}{l_m} \text{ and } H_g = \frac{M}{l_g} \tag{7}$$

The modified total flux, ϕ^*, is given by

$$\phi^* = H_g A_g + H_m A_m = I' A_m \tag{8}$$

The modified load line λ^* is then determined from the ratio I'/H_m, as follows:

$$-\lambda^* = \frac{I'}{H_m} = \frac{A_g}{A_m}\,\frac{l_m}{l_g} + 1 \tag{9}$$

This result, of course, is in agreement with equations 3 and 6. The same procedure is applied in the following paragraphs to a simple configuration.

Experimental Results

Experiments were conducted on the arrangement shown in Fig. 6 to verify the theoretical conclusions. Two pole pieces of mild steel were attached to a slotted ring magnet of barium ferrite. A rotating-coil type of field probe was then inserted through the gap. The field in this region should not differ from the field inside the magnet by more than a few per cent.

The external field was supplied by means of two sets of large ceramic ring magnets and the field was varied by changing the number of magnets on each side. Fig. 6 shows only the first two magnets, corresponding to the smallest external field. The demagnetizing field was measured with from zero to five magnets on each side. The small ring magnet was then replaced by copper spacers so that the exact values of the corresponding applied fields could be determined using the same field probe.

In order to prove the practical value of the static approach discussed above, the "modified load line" was computed by determining the field between two parallel finite pole faces and integrating it over the plane of symmetry to find the total flux. This calculation (given in the Appendix) leads to the following equation for λ^*:

$$\lambda^* = \frac{A}{A_m}\,\frac{1}{\kappa_0} \tag{10}$$

where κ_0 is a function of the ratio of the length of the pole pieces to the distance between them. When the numerical values $A_m = 2.0$ square centimeters, $A = 5.6$ square centimeters, and $\kappa_0 = 0.75$ (obtained from Fig. 9 for $\eta = 4$) are inserted, the load line is then given by $\lambda^* = 3.74$ ($\lambda = 2.74$).

The values given above are used in Fig. 7 together with a demagnetization curve for "Indox I" which was plotted at the David Sarnoff Research Center at Princeton. Because the measured values for H_m can be assigned to either the proper load line or the demagnetization curve, a "triangle of disagreement" is shown for both approaches, i.e., the modified load line with the intrinsic curve (solid lines), and the erroneous **B-H** construction (broken lines). It is quite obvious that only the intrinsic demagnetization curve yields correct values, and furthermore that the load line, computed above, is extremely accurate.

Five series of measurements were made with different dimensions and demagnetization curves and all produced the same convincing results. Naturally the question will be raised how such a discrepancy can have passed unnoticed to date. One reason is that the electrodynamic analogy with its lumped reluctances does not permit computation of the load line with any accuracy for more complicated configurations, and "stray factors" and "permeance factors" absorb any error. More important, however, is the fact that with older permanent-magnet materials (even including most of the Alnicos) the difference between the external and the

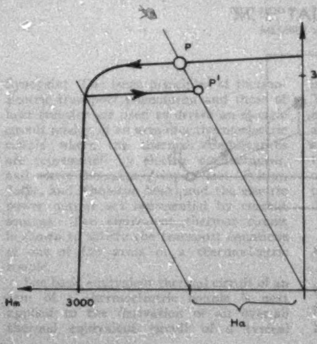


Fig. 4. Improved design construction, using the I'-H curve

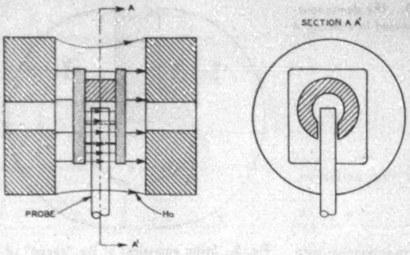


Fig. 6. Arrangement used to determine the effect of an external field on a magnetic circuit

intrinsic demagnetization curves is small. For a modern ceramic magnet, however, the intrinsic coercive force is usually even higher than the remanence.

Optimum Design

The question now arises whether this new knowledge will lead to different optimum design values. According to conventional practice, the load line should intersect the demagnetization curve at the point (**BH**) maximum for maximum effect per volume. It can be shown that this requirement is generally applicable for a closed system, in which the magnet has not been stabilized.[5]

Basically the concept of the optimum working point is not affected by the conclusions of this paper, because no external fields are involved. The process of "matching" the magnet to a given requirement, on the other hand, often involves modifications of the design and, therefore, of the useful field to be obtained. These modifications can lead to optimum results at working points different from the (**BH**) maximum point, especially if alien fields are involved. It is therefore advisable to examine each configuration in the same manner, as it has been done in the Appendix for the geometry shown in Fig. 6.

Conclusions

The observations discussed indicate that with modern magnetic materials the intrinsic magnetization curve is to be preferred for design purposes, especially when external fields are present. Furthermore, it appears mandatory to treat permanent-magnet problems as static-field configurations. This approach was used to analyze quite complicated field configurations in periodic focusing stacks for traveling-wave tubes with gratifying success.[6] A considerable number of magnet assemblies were designed on the basis of this analysis which produced fields within ±5% of the predicted values.

Appendix

The following procedure is used to compute the field component normal to the plane of symmetry for all values of z caused by two parallel plates which have given dimensions, as shown in Fig. 8, and are loaded with a uniformly distributed charge q:
The differential force in a unit pole located at z_1 is given by

$$\frac{1}{2}\,d^2H = q\,dx\,d\zeta\,\frac{1}{\rho^2}$$

where

$$\rho^2 = r^2 + x^2 = y^2 + (z-\zeta)^2 + x^2$$

The perpendicular component is given by

$$\frac{1}{2}\,d^2H_y = \frac{Y}{\rho}\,\frac{1}{2}\,d^2H = q\,Y\,dx\,d\zeta\,\frac{1}{\rho^3}$$

After integration, this component becomes

$$H_y = 4\int_{-R}^{+R}\int_0^{R/\sqrt{2}} q\,\frac{dx\,d\zeta}{[Y^2+(2-\zeta)^2+x^2]^{3/2}}$$

$$= 4q\left[\operatorname{arctg}\frac{R}{\sqrt{2}}\,\frac{z+R}{Y}\,\frac{1}{\sqrt{Y^2+\frac{1}{2}R^2+(z+R)^2}} - \operatorname{arctg}\frac{R}{\sqrt{2}}\,\frac{z-R}{Y}\,\frac{1}{\sqrt{Y^2+\frac{1}{2}R^2+(z-R)^2}}\right]$$

$$= 4q\kappa$$

Fig. 7 (right). The disagreement between computed and measured values (shaded areas) obtained by means of the conventional procedure (broken lines) and the corrected method (solid lines)

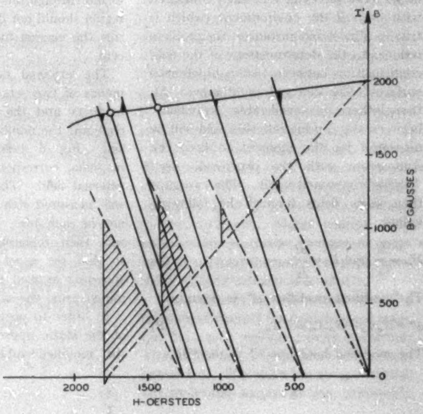
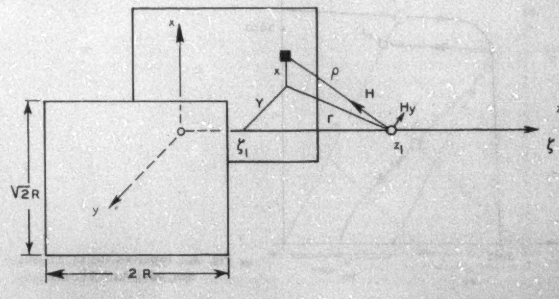


Fig. 8 (below). Computation of the field between two charged plates at point Z_1



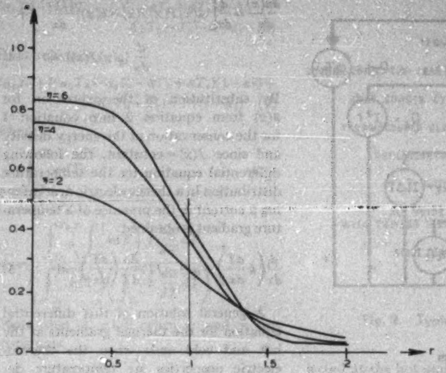


Fig. 9. The field distribution between flat pole pieces at different diameter-to-distance ratios

After substitution of the values $z=r.R$ and $R=\eta Y$, the field can be computed as function of the plate geometry, expressed by η, at a distance r from the origin, where $r=1$ is the point closest to the plate edges. The assumed surface charge q is now replaced by a quantity which can be attributed to the magnet. If the magnet cross section is not equal to the plate area, the surface charge on the magnet is given by

$$q_m = q\frac{A}{A_m}$$

where A denotes the plate area and A_m the magnet cross section. This surface charge q_m, however, is identical with the "intensity of magnetization" usually denoted as I. The intrinsic induction is equal to

$$4\pi I = 4\pi q\frac{A}{A_m} = I'$$

The field can then be expressed as

$$H = H_m = I'\frac{A_m}{A}\kappa$$

where κ is a function of geometry and location, as depicted in Fig. 9. The maximum (center) field increases with η and for infinitely big plates (or infinitely small spacing) would reach the value 1. Furthermore, the field starts to decrease about midway between the plate center and the edge, reaching half the center value in the edge region. However, measurements indicate a concentration of the charge in the edge region, which leads to a somewhat slower decrease of the field.

To determine the load line, it is necessary to know the total intrinsic flux ϕ^*. In the simple case of two parallel plates, the flux is given by $I'A_m$, as can easily be verified in the extreme cases of infinitely large plates ($\kappa=1$), where $\phi^*=HA=I'A_m$, or infinitely small plates, where the charge can be considered concentrated and $\phi^*=4\pi(Aq)=I'A_m$. Integration of the calculated field, plotted in Fig. 9, should of course yield the same result. This was found to be the case, with the greatest error of 10% occurring at $\eta=2$.

The expression for the load line is given by $\lambda^*=\phi^*/A_mH_m$. Because κ_0, the maximum value of κ is decisive to safeguard against local demagnetization, the following substitution can finally be made.

$$\lambda^* = \frac{1}{\kappa_0}\frac{A}{A_m}$$

References

1. Stabilized Permanent Magnets, P. P. Cioffi. *AIEE Transactions (Electrical Engineering)*, vol. 67, pt. II, 1948, pp. 1540–43.

2. New Method for the Optimum Design of Permanent Magnets Subjected to Demagnetizing Effects, Hans K. Ziegler. *Ibid.*, pt. I (*Communication and Electronics*), vol. 72, July 1953, pp. 253–63.

3. Analysis and Design of Permanent Magnet Assemblies, C. A. Maynard. *Machine Design*, Cleveland, Ohio, Apr. 18, 1957, pp. 122–43; see p. 137, "External Demagnetizing Influences."

4. Ferromagnetism (book), R. M. Bozorth. D. Van Nostrand Company, Inc., Princeton, N. J., 1951, p. 2.

5. The Economic Utilization of Modern Permanent Magnets, D. J. Desmond. *Journal of the Institute of Electrical Engineers*, London, England, vol. 92, 1945, pp. 229–52.

6. The Magnetic Field and Flux Distributions in a Periodic Focusing Stack for Traveling-Wave Tubes, M. J. Schindler. *RCA Review*, Princeton, N. J., vol. 21, Sept. 1960, pp. 414–36.

Electric Analog of Transport Phenomena in a Thermoelectric Generator

M. E. TALAAT
MEMBER AIEE

Synopsis: The basic principles of thermoelectric transport phenomena and those of heat transfer are used to derive an electric circuit analog of an arm of a thermoelectric couple where the thermal conductances are represented by electric conductances, and where the rates of input heat (Peltier, Joule, and Thomson heat) and the electric power output are represented by current sources. The equivalent thermal circuit is shown to satisfy the transport equations of one of the arms of a thermoelectric couple.

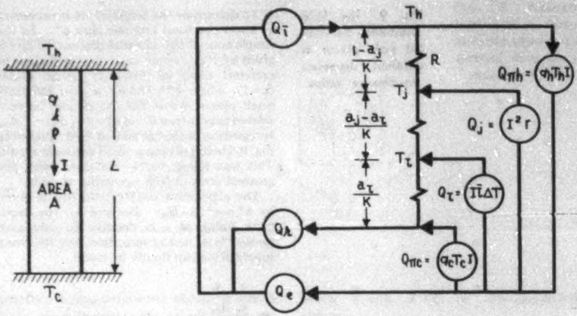
The basic equivalent thermal circuit of an arm of a thermoelectric couple is next applied to the derivation of an over-all thermal equivalent circuit of a typical isotope-fueled thermoelectric generator. The electric analog of the entire system enables the designer to determine the over-all efficiency of the thermoelectric generator as well as its temperature distribution, and to set the problem on the analog or digital computer and vary the different parameters of the circuit to optimize the over-all design.

THERMOELECTRIC generators with nuclear and other heat sources are being built today in sizes ranging from a few watts to several kw. The rate of heat generated by radioactive decay of radioisotopes is directly converted to electric power outputs ranging from a few to several hundred watts for auxiliary power sources in space vehicles and in all applications where small quantities of electric power are needed and remote or isolated operation over long time periods are necessary. For applications requiring electric power sources ranging from a few kilowatts to several hundred kilowatts, nuclear reactor heat sources can be used.

The over-all heat-to-electric energy conversion efficiencies which can be attained by these thermoelectric generator systems are dependent upon the following:

Paper 61-188, recommended by the AIEE Solid State Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 23, 1960; made available for printing December 2, 1960.

M. E. Talaat is with The Martin Company, Baltimore, Md.



1(A)

1(B)

Fig. 1. Electric circuit analog of a thermoelectric arm

1. Thermoelectric properties of the materials available and their useful temperature range.
2. Type of heat source used, the mechanism of heat transmission, and the corresponding hot junction temperature attainable.
3. The heat sink medium, whether outer space, environmental air, or sea water, and the corresponding limitations on size and weight of this heat sink.
4. The individual mechanical structural arrangement of the generator and the type of insulation and structural materials used.

To achieve optimum design considering simultaneously the thermoelectric materials, heat source, heat sink, and structural arrangement, a method of calculation is needed permitting evaluations of the over-all generator system and determination of the significant parameters, which can be modified to yield an optimum over-all design.

The method of calculation presented is intended to satisfy this need. It makes use of the concept of equivalent thermal circuits in which the power output from the heat source, the rate of Peltier, Joule, and Thomson heat, respectively, and the electric power output are taken to be analogous to electric current sources, and the inverse of thermal conductances are represented by their analogous electric resistances.

Basic Thermal Circuit of Thermoelectric Arm

Each arm of a thermoelectric couple is considered to be adiabatically insulated along its length with the hot end maintained at temperature, T_h, and the cold end maintained at temperature, T_c, and with a current, I, flowing through the arm from the hot end to the cold end for a P-type arm, and from the cold end to the hot end for an N-type arm. Then, by the Peltier effect, the arm absorbs heat at a rate equal to $\alpha_h T_h I$ watts at the hot junction at a temperature, T_h, and rejects heat at a rate equal to $\alpha_c T_c I$ watts at the cold junction at a temperature, T_c. Also, as a result of thermal conduction, the arm absorbs heat at a rate equal to $(-Ak_h \nabla T_h)$ watts at the hot junction and rejects heat at a rate equal to $(-Ak_c \nabla T_c)$ at the cold junction.

Now, the rate of Peltier heat absorbed or evolved can be readily evaluated from the temperatures, T_h and T_c, and the corresponding Seebeck emf (electromotive-force) coefficients, α_h and α_c, at the hot and cold ends, respectively. However, the rate of heat absorbed or rejected by the arm as a result of thermal conduction involves the temperature gradient at the hot and cold ends and thus requires the solution of the differential equation for the temperature distribution in a thermoelectric arm carrying a current in the presence of a temperature gradient.

This differential equation is derived from consideration of the transport equations for the heat current density, $q(x)$, and the electric current density, $J(x)$, as given by

$$q(x)=\alpha(x)T(x)J(x)-k(x)\frac{dT(x)}{dx} \qquad (1)$$

$$J(x)=-\sigma(x)\left[\frac{d\bar{\mu}(x)}{dx}+\alpha(x)\frac{dT(x)}{dx}\right] \qquad (2)$$

Now, the total energy flux density $u(x)$ is given by

$$u(x)=q(x)+J(x)\bar{\mu}(x) \qquad (3)$$

and since, for an arm which is adiabatically insulated along its length, the energy density $u(x)$ is conserved, i.e., $du(x)/(dx)=0$,

$$\frac{du(x)}{dx}=\frac{d}{dx}\left[\alpha(x)T(x)J(x)-k(x)\frac{dT(x)}{dx}\right]+\frac{d}{dx}\,[\bar{\mu}(x)J(x)]=0 \qquad (4)$$

By substitution of the expression for $\bar{\mu}(x)$ from equation 2 into equation 4 for the conservation of the energy density and since $J(x)=\text{constant}$, the following differential equation for the temperature distribution in a thermoelectric arm carrying a current in the presence of a temperature gradient is obtained:

$$\frac{d}{dx}\left(k\frac{dT}{dx}\right)+\rho J^2-JT\frac{d\alpha}{dT}\left(\frac{dT}{dx}\right)=0 \qquad (5)$$

A general solution of this differential equation for the thermal gradients at the hot and cold ends when the thermoelectric properties are temperature dependent was derived by Burshtein.[1] He first considered the influence of the electric effects alone on the thermal conduction process and assumed then that the interaction of the electric effects involves higher order terms which may be neglected. Burshtein's approximate general solutions for the rate of heat absorbed at the hot end and the rate of heat rejected at the cold end by the thermal conduction process is given by

$$-Ak_h\left(\frac{dT}{dx}\right)_h=\frac{A}{L}\int_{T_c}^{T_h}kdT-I\frac{\int_{T_c}^{T_h}\left(\int_{T_c}^{T}kdT\right)T\frac{d\alpha}{dT}dT}{\left(\int_{T_c}^{T_h}kdT\right)}-I^2\frac{L}{A}\frac{\int_{T_c}^{T_h}\left(\int_{T_c}^{T}kdT\right)\rho kdT}{\left(\int_{T_c}^{T_h}kdT\right)^2} \qquad (6)$$

$$-Ak_c\left(\frac{dT}{dx}\right)_c=\frac{A}{L}\int_{T_c}^{T_h}kdT+I\frac{\int_{T_c}^{T_h}\left(\int_{T_c}^{T}T\frac{d\alpha}{dT}dT\right)kdT}{\left(\int_{T_c}^{T_h}kdT\right)}+I^2\frac{L}{A}\frac{\int_{T_c}^{T_h}\left(\int_{T_c}^{T}\rho kdT\right)kdT}{\left(\int_{T_c}^{T_h}kdT\right)^2} \qquad (7)$$

By using Burshtein's solution and by including the rate of Peltier heat absorbed or rejected, the following expressions are obtained for the total rate of heat absorbed or rejected at the hot and cold ends, respectively:

Heat absorbed:

$$I\alpha_h T_h - I(\alpha_h T_h - \alpha_c T_c - \bar{\alpha} T_h + \bar{\alpha} T_c) a_\tau + K(T_h - T_c) - I^2 r a_j \quad (8)$$

Heat rejected:

$$I\alpha_c T_c + I(\alpha_h T_h - \alpha_c T_c - \bar{\alpha} T_h + \bar{\alpha} T_c)(1 - a_\tau) + K(T_h - T_c) + I^2 r(1 - a_j) \quad (9)$$

where

$$\bar{\alpha} = \frac{\int_{T_c}^{T_h} \alpha\, dT}{T_h - T_c} \quad (10)$$

$$a_\tau = \frac{\int_{T_c}^{T_h} \left(\int_{T_c}^{T} k\, dT \right) T \frac{d\alpha}{dT} dT}{\left(\int_{T_c}^{T_h} k\, dT \right) \left(\int_{T_c}^{T_h} T \frac{d\alpha}{dT} dT \right)} \quad (11)$$

Note that $a_\tau = 1/2$ when k and τ are constants.

$$K = \frac{A}{L} \frac{\int_{T_c}^{T_h} k\, dT}{T_h - T_c} = \frac{A}{L} \bar{k} \quad (12)$$

$$r = \frac{L}{A} \frac{\int_{T_c}^{T_h} k\rho\, dT}{\int_{T_c}^{T_h} k\, dT} = \frac{L}{A} \bar{\rho} \quad (13)$$

$$a_j = \frac{\int_{T_c}^{T_h} \left(\int_{T_c}^{T} k\, dT \right) k\rho\, dT}{\left(\int_{T_c}^{T_h} k\, dT \right) \left(\int_{T_c}^{T_h} k\rho\, dT \right)} \quad (14)$$

Note that $a_j = 1/2$ when k and ρ are constants.

The equivalent thermal circuit for an arm of a thermoelectric couple which satisfies the equations for the rate of total heat absorbed, equation 8, and the rate of total heat rejected, equation 9, at the hot and cold ends, respectively, is presented in Fig. 1(B). To prove that this is the case, the node equations for the circuit given in Fig. 1 are written as follows:

$$Q_t = Q_{\pi h} + \frac{T_h - T_j}{1 - a_j} K \quad (15)$$

$$\frac{T_j - T_\tau}{a_j - a_\tau} K = \frac{T_h - T_j}{1 - a_j} K + Q_j \quad (16)$$

$$\frac{T_\tau - T_c}{a_\tau} K = \frac{T_j - T_\tau}{a_j - a_\tau} K + Q_\tau \quad (17)$$

$$Q_r = \frac{T_\tau - T_c}{a_\tau} + Q_{\pi c} \quad (18)$$

By solving for T_j and T_τ from equations 16 and 17, and inserting their values in the expressions for Q_t (equation 15) and Q_r (equation 18), Q_t and Q_r as given by equations 8 and 9 are obtained.

Thus, it can be concluded that the equivalent circuit presented here satisfies the heat current equations for the rate of heat absorbed and the rate of heat rejected at the hot and cold ends, which are obtained from the solution of the differential equation for the temperature distribution in a thermoelectric arm carrying a current in the presence of a temperature gradient, equation 5. The latter in turn is derived from the heat current and electric current transport equations for an arm of a thermoelectric couple, equations 1 and 2.

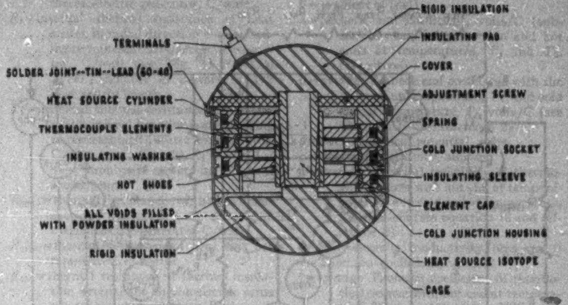


Fig. 2. Typical arrangement of isotope-fueled thermoelectric generator

Application of Basic Equivalent Thermal Circuit to Typical Thermoelectric Generators

A typical arrangement of an isotope-fueled thermoelectric generator will now be considered; see Fig. 2. The heat generated by the radioactive decay of the encased radioisotope flows radially across the contact film between the isotope core and the inner casing, the inner casing shell, a second contact film between the inner casing and an outer casing, the outer casing shell, the contact film between the outer casing and the heat cylinder, the heat cylinder shell, the electrical insulation between the heat cylinder and the hot junction bridges, the hot junction bridges, the thermoelectric arms and the insulation surrounding them, the cold junction bridges, the insulation between the cold junction bridges and the heat sink, and finally across the heat sink to the cooling medium. Heat also flows axially across the film between the isotope core and the casing, the casing shell, the thermal insulation, and finally across the axial heat sink portion to the cooling medium.

Because the thermoelectric couples in generator designs are usually symmetrically arranged around the heat source, one has only to consider a representative sector of that arrangement and derive the equivalent thermal circuit for that sector using the basic equivalent thermal circuit given in Fig. 1 for each thermoelectric arm. Fig. 3 gives the over-all equivalent thermal circuit of the thermoelectric generator arrangement shown in Fig. 2. In this over-all equivalent thermal circuit, Q_t is the current source analog of the rate of heat generated by that part of the heat source corresponding to the representative sector. R_t and R_t' are the corresponding radial and axial internal thermal resistances of the nuclear heat source. In this circuit representation, the current source analogs, Q_{hj} and Q_{cj}, of the rate of Joule heat, due to the hot and cold junction resistances, are also shown.

Application of Circuit to Design of Thermoelectric Generators

The equivalent thermal circuit representation given in Fig. 3 can be applied to the over-all design problem of a thermoelectric generator system. Two examples of such an application are discussed in the following.

SOLUTION WITH SPECIFIED HOT AND COLD JUNCTION TEMPERATURES

One solution would be to fix the hot and cold junction temperatures, T_h and T_c, to utilize to the maximum the capability of a given set of thermoelectric materials over their best temperature ranges, and then to determine the numbers and geometry of the thermoelectric arms to yield the prescribed optimum electrical performance. Next, calculate the equivalent circuit of the generator system, given in Fig. 3 in various mechanical arrangements, to evaluate the temperature distribution inside the generator and to determine the heat source power and/or heat sink rejection rate which is necessary to maintain the chosen hot and cold junction temperatures.

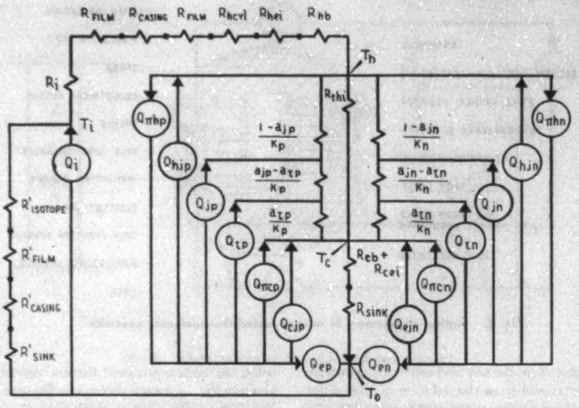


Fig. 3. Typical electric circuit analog of isotope-fueled thermoelectric generator

In reference 2, a synthesis procedure is outlined for determining the number and geometry of segmented or nonsegmented thermoelectric generator arms for the condition of optimum efficiency or maximum power output once the temperature dependent thermoelectric properties of materials are known and the hot and cold junction temperatures are specified. In this paper, the equivalent thermal circuit of the over-all generator system presented in Fig. 3 may be used to determine the necessary requirements to maintain these specified hot and cold junction temperatures.

SOLUTION WITH HEAT SINK DESIGN AND TEMPERATURE FIXED BY SIZE AND WEIGHT LIMITATIONS

Another important case where the equivalent thermal circuit can be usefully applied is one where the heat sink design and operating temperatures are set by limitations on its size, weight, and cooling medium to be used. In this case, first assume the cold junction temperature, T_c, from consideration of the selected heat sink operating temperature and choose the hot junction temperature, T_h, from consideration of the hot temperature capability of the thermoelectric materials available. Next use the synthesis procedure outlined in reference 2 to determine the number and geometry of the thermoelectric arms corresponding to the assumed hot and cold junction temperatures which would yield the specified electrical output either for the condition of optimum efficiency or maximum power. Following this, calculate the equivalent thermal circuit parameters of the thermoelectric arms, insert their values in the equivalent thermal circuit of the generator system, calculate the temperature distribution and repeat this until the calculated heat sink temperature obtained from the equivalent circuit agrees with the heat sink design temperature. The resulting heat source current and its internal temperature determine the design requirements for the heat source.

Calculations using the equivalent circuit can be made either by using the analog computer or by writing the node equations and employing a digital computer for the reiterative procedure of calculation. The method described can be applied to other cases where the equivalent circuit method may be used.

Discussion

An important advantage of the equivalent thermal circuit method is that it provides a physical picture by which one can readily see all the parameters involved in the generator system. It also indicates their relative importance in influencing the efficiency of the generator, the temperature distribution inside the generator, and in determination of requirements on the heat source and heat sink. The relative importance of the various heat losses across the parts of the generator system, other than the thermal conduction losses involved in the heat-to-electric energy conversion process in the thermoelectric couples themselves, can also be determined. Any of these parameters may be modified to reduce these other losses or improve the temperature distribution. Needless to say, the predetermination of the generator temperature distribution is important if the problem of differential expansion and the resulting stresses inside the generator is considered, and if safe insulation and structural material temperatures are to be realized.

Finally, it must be mentioned that the equivalent thermal circuit of the thermoelectric generator presented here can also be used to match the heat flux from any other source of heat to the requirements of the thermoelectric energy conversion process. The heat transfer mechanism from the heat source to the heat cylinder or hot junction surface does not have to take place by conduction, as in the given example, but can take place by forced convection or radiation.

Conclusions

To summarize, a thermal equivalent circuit method by which the designer can obtain the optimum design arrangement for a thermoelectric generator has been established. The method also yields the dimensions and physical constants which enter into the expressions for the parameters of that thermal equivalent circuit. Therefore the designer can readily see the factors at his disposal that can reduce the power losses or optimize his design. It is hoped that this paper, together with reference 2, will help the thermoelectric generator designer to control the performance of the thermoelectric generator while in the design stages and, thus, to attain the prescribed performance without recourse to curative measures which may slow production and raise costs.

Nomenclature

(For more detailed definition of symbols, see reference 2.)

A = area of thermoelectric arm, cm^2 (centimeter^2)
a_i = radius of heat source cylinder
a_j = fraction of Joule heat which is fed back at hot junction
a_τ = fraction of Thomson heat which is fed back at hot junction
c = subscript referring to cold junction
d = letter used to denote differentiation
h = subscript referring to hot junction
I = current through thermoelectric arm, amperes
J = current density through thermoelectric arm, amperes/cm^2
K = thermal conductance of the thermoelectric arm with the hot end at temperature, T_h, and cold end at temperature, T_c, watts/C (degree centigrade)
k = thermal conductivity, watts/C–cm
L = total length of thermoelectric arm, cm
L_i = length of heat source cylinder

n = subscript referring to N-type thermoelectric arm
p = subscript referring to P-type thermoelectric arm
*Q_{cj} = rate of Joule heat due to cold junction resistance, watts (see reference 2)
*Q_e = electric power output, watts (see reference 2)
*Q_{hj} = rate of Joule heat due to hot junction resistance, watts (see reference 2)
*Q_i = rate of input heat at hot junction, watts
*$Q_j = I^2r$ = rate of Joule heat generated inside thermoelectric arm with hot end at temperature, T_h, and cold end at temperature, T_c, watts
*Q_r = rate of rejected heat at cold junction, watts
*$Q_{\pi h} = \alpha_h T_h I$ = rate of Peltier heat absorbed at hot junction, watts
*$Q_{\pi c} = \alpha_c T_c I$ = rate of Peltier heat rejected at cold junction, watts
*$Q_\tau = I\bar{\tau}\Delta T = I\bar{\tau}(T_h - T_c)$ = rate of Thomson heat generated inside thermoelectric arm with hot end at temperature, T_h, and cold junction at temperature, T_c, watts
q = heat current density, watts/cm^2 (see equation 1)
R_{cb}, R_{hb} = thermal resistance of cold and hot junction bridges, respectively, per representative sector of the thermoelectric generator, C/watt
R_{cei}, R_{hei} = thermal resistance of electrical insulation at cold and hot ends, respectively per representative sector of thermoelectric generator, C/watt
R_i = internal thermal resistance of heat source in radial direction corresponding to representative sector of generator, C/watt (see Fig. 3) $R_i = 1/k_i\; 2/4\pi L_i$ number of sectors in 1/2 heat source length
R_i' = internal thermal resistance of the heat source in axial direction corresponding to representative sector of generator, C/watt (see Fig. 3) $R_i' = 1/k_i\; L_i/4\pi a_i^2$ number of sectors in 1/2 heat source length
R_{hycl} = thermal resistance of hot cylinder shell per representative sector of the thermoelectric generator C/watt
R_{sink} = thermal resistance of heat sink per representative sector
R_{thi} = thermal resistance of thermal insulation around the thermoelectric arms per representative sector of thermoelectric generator, C/watt
r = total electrical resistance of thermoelectric arm with hot end at temperature, T_h, and cold end at tempeature, T_c, ohms
T_o = cooling medium temperature, K (degrees Kelvin)
T_c = cold end temperature, K
T_h = hot end temperature, K
T_j = see equations 15–17
T_τ = see equations 15–18
u = total energy flux density, watts/cm^2 (see equation 3)
x = distance along thermoelectric arm measured from hot end
∇ = gradient of (del)
α = Seebeck emf coefficient, volts/C (subscripts c and h denote cold and hot ends at temperatures, T_c and T_h, respectively)
$\bar{\alpha}$ = average Seebeck emf coefficient with the not end at temperature, T_h, and cold end at temperature, T_c, volts/C (see equation 10 and reference 2)
$\bar{\mu}$ = electrochemical potential, volts
ρ = electrical resistivity, ohm-cm
$\bar{\rho}$ = average electrical resistivity of thermoelectric arm with hot end at temperature, T_h, and cold end at temperature, T_c, ohm-cm (see equation 13 and reference 2)
$\sigma = 1/\rho$ = electrical conductivity, (ohm–cm)$^{-1}$
τ = Thomson coefficient, volts/C
$\bar{\tau}$ = average Thomson coefficient of thermoelectric arm with hot end at temperature, T_h, and cold end at temperature, T_c, volts/C
$\bar{\tau} = (\alpha_h T_h - \alpha_c T_c - \bar{\alpha} T_h - \bar{\alpha} T_c)/(T_h - T_c)$

References

1. An Investigation of the Steady-State Heat Flow Through a Current Carrying Conductor, A. I. Burshtein. *Soviet Physics JETP*, New York, N. Y., vol. 2, no. 7, 1958, p. 1397.

2. Synthesis of Thermoelectric Generators, M. E. Talaat. *Electrical Engineering*, vol. 79, no. 6, July 1960, pp. 476–81.

---

* In Fig. 3, Q_J' should be those corresponding to the number of arms per representative sector of the thermoelectric generator.

On the Process of Flux Reversal in Multiaperture Ferrite Cores

IAN H. ROWE
NONMEMBER AIEE

GORDON R. SLEMON
MEMBER AIEE

TOROIDAL MAGNETIC cores using ferrite materials have been applied extensively as 2-state logic and memory devices.[1–4] For more complex logical operation, multiaperture ferrite cores have been used.[5–8] In most of the studies of flux reversal phenomena in these cores only two final flux states have been considered, representing either of the two saturated states of the core or a leg of the core. Multiaperture cores are also capable of continuous modulation of a-c power. An example is the transfluxor,[9–11] a 2-aperture core, in which the magnitude of the output is controlled by a transient setting pulse applied to its control winding. This device is able to operate at a continuously-controllable level because the amount of magnetic flux available about one aperture for the transfer of power can be controlled by the particular magnetic state established about the other aperture.

In order to devise, on a rational basis, multiaperture configurations for a variety of continuous-control functions it is necessary to understand the manner in which the magnetic state of each part of the ferrite core is influenced by the winding currents acting either individually or collectively. Abbas[12, 13] has calculated potential and stream functions for some core configurations by assessing certain distributions of magnetic poles on the core surfaces and flux reversal boundaries.

The purposes of this paper are (a) to present some experimental results obtained with 2-aperture cores, and (b) to develop a model representing the flux states within a 2-aperture core for various winding current combinations. It is expected that the general principles used in modelling a 2-aperture core will be applicable for all multiaperture ferrite cores.

Flux Reversal in Disc With Single-Aperture

In order to demonstrate some of the general principles governing the flux state of a ferrite core let us consider first two elementary toroidal discs each with a single centered aperture as shown in Fig. 1(A) and (B). Measured relationships for the magnetic flux ϕ in the disc as a function of the mmf (magnetomotive force) F established by a winding through the aperture are shown in Fig. 1(C) and (D) for the discs of Fig. 1(A) and (B). It is noted that the ϕ/F characteristics are reasonably linear over most of the excur-

---

Paper 61-128, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1960. Manuscript submitted August 12, 1960; made available for printing December 28, 1960.

Ian H. Rowe is with De Havilland Aircraft of Canada, Ltd., Downsview, Ont., Canada, and Gordon R. Slemon is with the University of Toronto, Toronto, Ont., Canada.

The research on which this paper is based was supported by a grant from the Defense Research Board to the Department of Electrical Engineering, University of Toronto.

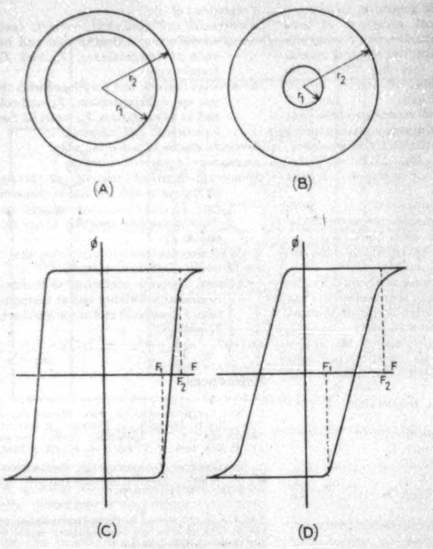


Fig. 1. Ferrite discs having apertures of different radii (A, B), magnetic flux ϕ plotted for disc in (A) on magnetomotive force F (C), and as in (C) for the (B)-disc, (D)

sion between saturated states. If the curved end regions or fillets are neglected for the moment and the linear portion is extended to intercept the saturation flux lines at F_1 and F_2 it is found that, within reasonable accuracy,

$$\frac{F_2}{F_1}=\frac{r_2}{r_1} \tag{1}$$

where r_1 and r_2 are the inner and outer disc radii.

This experiment suggests that a disc with a single centered aperture may be considered as a series of independent concentric annuli and that the basic property of the ferrite material may be represented by the essentially-rectangular **B/H** characteristic shown in Fig. 2. An idealized model of the flux state of a disc is shown in Fig. 3. This disc was originally saturated in a counterclockwise direction and then had a clockwise magnetomotive force, F, applied. The radius, r, of the flux reversal boundary is related to F by

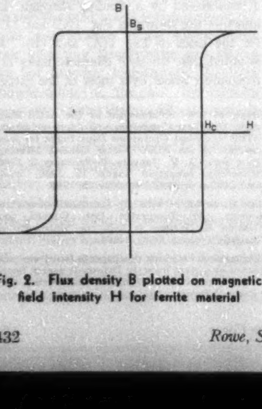


Fig. 2. Flux density B plotted on magnetic field intensity H for ferrite material

$$F=H_c\cdot 2\pi r \tag{2}$$

where H_c is the coercive force in Fig. 2.

Although this model may appear obvious, it is by no means the only one which could have been proposed to explain the results of Fig. 1. As a test of this model consider the effect of applying an alternating mmf to the core. Equation 2 suggests that, during the first half-cycle, flux reversal will commence at the aperture periphery and a ring of reversed magnetization will grow outward. On subsequent reversals of the applied mmf new rings of reversed magnetization, each in a direction opposite to its predecessor, will grow outward from the inner surface of the core. Under no circumstances can a flux reversal boundary be made to move inward. These statements are supported experimentally as described in the following.

A small hole was drilled through a disc so that inner and outer search windings through the hole enclosed equal cross-sectional areas of the core as shown in Fig. 4. A sinusoidal voltage source was used to drive the core.[14,15] Fig. 5 shows oscillograms of the induced voltages in the inner and outer search coils for various degrees of core excitation. It is noted that an induced voltage (and therefore flux change) occurs in the inner coil until the flux reversal boundary passes the sense hole. Then, most of the remaining portion of the applied sine wave is reproduced in the outer coil.

The flux-state model of Fig. 3 is an idealization of an actual core. In ferrites there are many local variations in the material caused by voids, occlusions, and stress conditions. The sintering process assures that these are randomly distributed. Thus, while the core material is highly disordered when viewed microscopically, it assumes the nature of a nearly homogeneous material when viewed macroscopically. The shape of the **B/H** characteristic of Fig. 2 may be taken as evidence of the high degree of homogeneity in the material. A large proportion of the domains in any unit volume of the material experiences flux reversal at essentially the same value of field intensity H_c. For these, the energy input per unit volume to produce reversal is $B_s\cdot H_c$. A very small proportion experiences reversal for $H<H_c$. The fillet in the **B/H** characteristic, as saturation density B_s is approached, is a measure of the proportion of a unit volume requiring an energy per unit volume greater than $B_s\cdot H_c$ for reversal. Small regions requiring $H>H_c$ for reversal appear to be left behind as the main flux reversal boundary moves across the core. Examination of the waveforms in Fig. 5 shows that the voltage wave on the inner search winding has a significant decaying tail. This is partially the result of the process of clearing this inner section of small regions requiring higher-than-average reversal energy. The voltage wave on the outer search

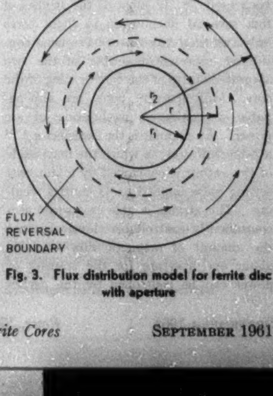


Fig. 3. Flux distribution model for ferrite disc with aperture

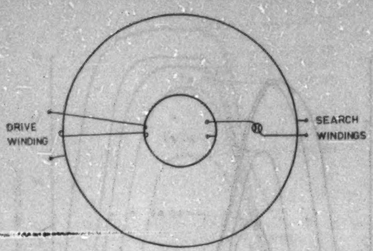


Fig. 4. Ferrite disc with sense hole for search coils

winding has a reasonably abrupt rise which may be considered a result of the main flux reversal boundary crossing the sense hole. Of course, the sense hole itself causes a local inhomogeneity and may therefore cause some of the imperfection in the transition from the voltage on the inner winding to the voltage on the outer winding.

The fact that the relative saturated permeability is greater than unity is normally attributed to domain rotation, a process in which magnetic stored energy is essentially conserved. In Fig. 5 this effect is indicated by a small contribution of induced voltage simultaneously in both inner and outer search windings at the beginning of each half-cycle of applied voltage. These contributions occur during the rapid reversal of the core mmf and are in a direction opposite to that of the previously induced voltage.

From the previous discussion it appears that the flux state of a disc with a single centered aperture is governed by the following general principles:

1. Essentially all parts of the core are at saturation flux density B_s.
2. Magnetic flux is established in continuous closed paths within the core.
3. An energy of approximately $B_s \cdot H_c$ per unit volume is required to move a flux reversal boundary outward from an aperture.

Let us now consider the application of these principles in the determination of the flux state of a 2-aperture core under various conditions of operation.

Experiments on Disc With Two Apertures

A number of experiments were performed on a 2-aperture "transfluxor" core shown in Fig. 6. The two apertures are of unequal diameter and form three legs, numbered 1, 2, and 3. The cross-sectional area of leg 1 is slightly greater than the sum of the equal areas of legs 2 and 3. The core was made from a magnesium-manganese (MgMn) ferrite material.

The core was first examined under a set of conditions that will be termed the Set-Drive mode of operation. The control winding W_1 was connected to a current source that could supply an amplitude-controlled long-duration pulse of either sign. In the manner described by Rajchman, et al,[9,10] the core was completely saturated in one direction (blocked) and then a setting mmf was applied in the opposite direction. The winding W_2 was then connected to a sinusoidal current drive at 60 cps (cycles per second). The voltage on winding W_3 provided a measure of the flux in volt-seconds that was being switched about the smaller aperture.

The setting characteristics for the Set-Drive mode in Fig. 7 were obtained in the following manner. For each point on the curves, the core was first blocked, then set with an appropriate current pulse in winding W_1 and then driven from the sinusoidal current supply connected to winding W_2. On extrapolating the curves in Fig. 7 it is noted that they all intersect the abscissa near a point F_0. Furthermore, the curves show a peak output at the mmf denoted as F_p. Setting characteristics were obtained from the Set-Drive mode with the current supply connected to winding W_3, and with the output measured on winding W_2. These were found to be essentially the same as those in Fig. 7.

Other modes of operation, not previously reported in the literature, are found when applying the winding currents in a different sequence. The Drive-Set mode was obtained in the following manner. The sinusoidal current drive was applied to winding W_3. A setting pulse lasting several cycles was applied to winding W_1 while the drive current was present. The setting characteristics for this mode are shown in Fig. 8. These curves can best be interpreted by replotting them as in Fig. 9 on a new abscissa which is the setting mmf F_S plus the peak magnitude of the drive mmf F_D. The dominant feature of the linear control region in this mode is that the output is linearly dependent on a quantity C_ϕ where

$$C_\phi = F_S + F_D \tag{3}$$

Although the set and drive currents do not pass through the same aperture, their magnetomotive forces add linearly to establish the magnetic state of the core. The experiment of the Drive-Set mode was repeated with the drive connected to winding W_2 instead of winding W_3. This connection assured that the closed magnetic path around the periphery of the larger aperture contained both the set mmf and the drive mmf. The results were found to be essentially the same as those shown in Figs. 8 and 9.

To examine further the interaction of

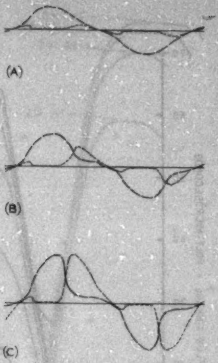


Fig. 5. Waveforms of voltages induced in inner and outer search coils of Fig. 4

A—40% excitation
B—60% excitation
C—90% excitation

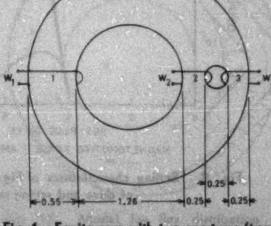


Fig. 6. Ferrite core with two apertures (transfluxor) showing dimensions in centimeters

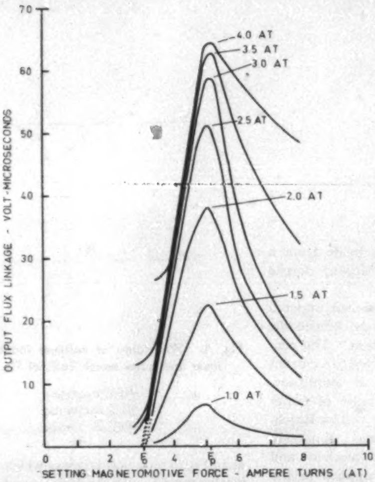


Fig. 7. Setting characteristics for Set-Drive mode. Drive—leg 2; output—leg 3

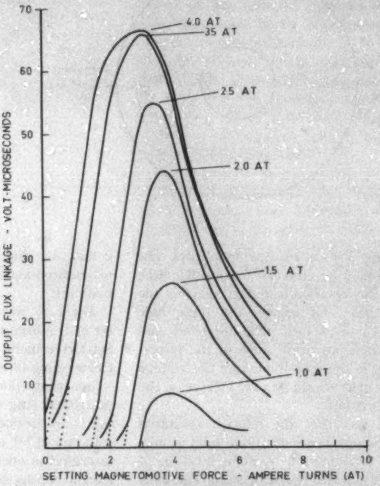


Fig. 8. Setting characteristics for Drive-Set mode. Drive—leg 3; output—leg 2

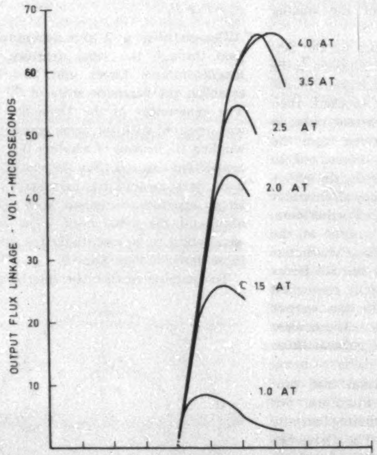


Fig. 9. Setting characteristics of Fig. 8 replotted on sum of drive and setting mmfs

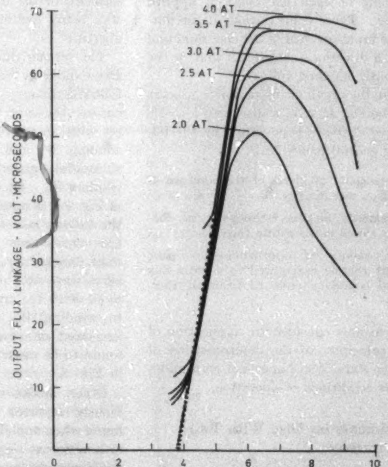


Fig. 10. Setting characteristics for Coincident-Drive-and-Set mode. Drive—leg 2; output—leg 3. Drive and setting mmfs opposing in leg 2

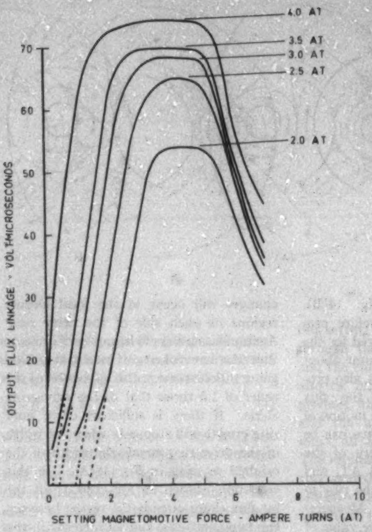


Fig. 11 (left). Setting characteristics for Coincident-Drive-and-Set mode. Drive—leg 2, output—leg 3. Drive and setting mmfs assisting in leg 2

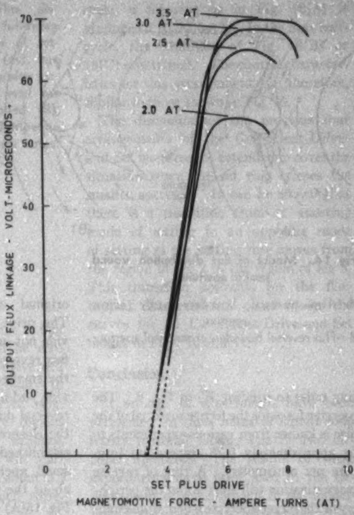


Fig. 12 (right). Setting characteristics of Fig. 11 replotted on sum of drive and setting mmfs

the set and drive mmfs acting together a third mode termed the Coincident-Set-and-Drive mode was investigated. The set pulse was made to coincide with the drive for one half-cycle only of the drive. A gating circuit was devised to admit a single half-cycle pulse from a sinusoidal current source to winding W_1 in coincidence with the first half-cycle of current gated through the drive winding. This method of gating assured that F_S and F_D remained proportional to each other during the half-cycle setting period. In the first instance the drive was applied to winding W_2. During the setting period the mmfs F_S and F_D were in opposition in the path around the larger aperture. The experimental results are shown in Fig. 10. They bear a distinct resemblance to Set-Drive characteristics shown in Fig. 7, the main difference being the behaviour at large values of setting mmf. A repetition of the experiment with the drive on winding W_3 gave essentially the same results as in Fig. 10 where the mmfs F_S and F_D were in opposition in the path around the core periphery during the setting period.

In the second instance of the Coincident-Set-and-Drive mode, the drive was applied in such a direction that the mmfs F_S and F_D were assisting in a path enclosing both during the setting period. The setting characteristics for this mode are shown in Fig. 11 for the drive on winding W_2. Essentially the same results were obtained with the drive on winding W_3. These characteristics are seen to be similar to those obtained for the Drive-Set mode of operation (Fig. 8). When replotted on an abscissa (F_S+F_D) as in Fig. 12, the linear dependence of output on the sum of the drive and setting mmfs is again obtained. The major points of difference between the characteristics of Fig. 12 and those of the Drive-Set mode in Fig. 9 are the relatively flat tops of the curves in Fig. 12.

Flux Model for 2-Aperture Disc

Here a model for the flux distribution within a 2-aperture disc is developed, using the general principles stated earlier. An attempt is made to show that this model is consistent with the experimental evidence and can be used for the approximate prediction of the behaviour of the core in any of its modes of operation.

First, consider the flux distribution which follows the blocking of the core in Fig. 6. On termination of the blocking pulse the core will relax magnetically into the lowest available energy state. One characteristic of this low-energy state will be that each element of magnetic flux, which must be continuous, will generally close within the core rather than through any air path. Fig. 13 shows a possible model for the 2-aperture core in the blocked state. To preserve flux continuity about the larger aperture, isolated regions on either side of the smaller aperture are left in a state of local closure. There is no direct experimental information on the nature of these local closure regions but fortunately such information is not necessary for the discussion to follow. In addition to these two closure regions there will be other regions of local closure of flux near the outer edge of the core because the saturation flux around the core is limited by its minimum cross-sectional area.

Then consider the application of a set-

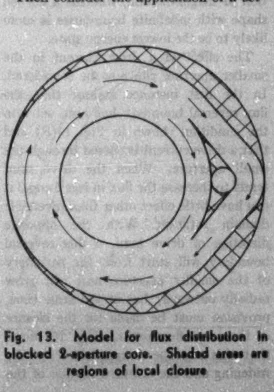


Fig. 13. Model for flux distribution in blocked 2-aperture core. Shaded areas are regions of local closure

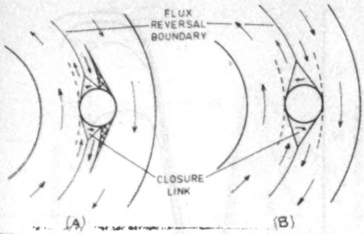


Fig. 14. Model of flux distribution around smaller aperture

A—Flux reversal boundary partly across aperture
B—Flux reversal boundary at center of aperture

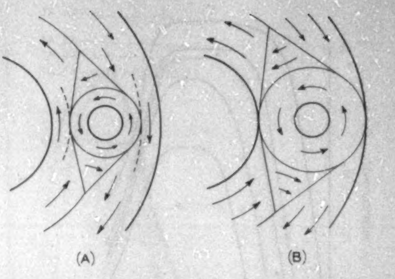


Fig. 15 (right). Model showing drive ring of flux around smaller aperture and closure links for core which is initially set as in Fig. 14(B)

A—Partial drive
B—Full drive for maximum output

ting pulse to winding W_1 in Fig. 6. The coercive force for the ferrite material of the disc is known from various experiments to be approximately 0.75 ampere/cm (ampere per centimeter). A ring of reverse magnetization will start about the periphery of the larger aperture when, from equation 2, the setting mmf F_D equals 3.0 AT (ampere-turns). This value is designated as F_0. The radius of the flux reversal boundary will increase linearly with $(F_D - F_0)$ until the boundary reaches and attempts to cross the smaller aperture. Fig. 14 shows one possible manner in which this boundary may cross the aperture region. Flux continuity is preserved by the absorption of a part or all of the local closure regions of Fig. 13 into the closure links which allow part of the set ring flux to close on the original, opposite-directed flux on the outer side of the boundary. For the sake of simplicity the closure links in Fig. 14 are shown with sharp boundaries and a single direction of magnetization, whereas a curved shape with indefinite boundaries is more likely to be the lowest energy state.

The effect of a drive current in the smaller aperture can now be considered. In the first instance assume that the flux reversal boundary has been set into the condition shown in Fig. 14(B) and that a drive current is passed through the small aperture. When the drive mmf tends to increase the flux in legs 2 and 3 it can have little effect other than reversible domain rotation. With the opposite direction of drive mmf, a flux reversal boundary will start from the periphery of the smaller aperture and will grow radially outward; but at the same time, provision must be made for the closure of the main flux paths. Fig. 15 shows that this may be accomplished by the widening and the outward motion of the original closure links in Fig. 14(B). The drive winding must therefore provide not only the energy required for the flux reversal in the circular volume about the smaller aperture, but must also provide additional energy for the flux reversal during the growth and motion of the closure links. Coercive force can be established about the periphery of the small aperture with $F_D = 0.59$ AT, and about the outer periphery of the ring in Fig. 15(A) with $F_D = 1.08$ AT. Measurements from Fig. 7 show that the value of drive mmf necessary to start ring growth is about 0.85 AT and that the value to produce the state of Fig. 15(A) is about 1.8 AT. These results suggest that the volume of material undergoing flux reversal may be about 1.5 times the ring volume.

If the drive mmf is alternating, the ring of magnetization established in the first cycle will be reversed on each subsequent half-cycle as with a single-aperture core. The output in Fig. 7 is, therefore, a direct measure of the extent of ring growth. With a sufficient value of drive mmf and starting from the original set condition of Fig. 14(B) the ring will grow to include all the width of legs 2 and 3. In this condition, shown in Fig. 15(B), all of the flux surrounding the larger aperture closes on itself in large closure links.

Now consider the manner in which the drive produces a flux ring when the core is initially in a partially set condition such as that of Fig. 16(A). Ring growth can take place only when the drive mmf is in such a direction as to reduce the flux in leg 3. As ring growth takes place [Fig. 16(B)] the preservation of flux continuity within the core requires a reversal of a segment of the original set ring in Fig. 16(A). The width of this segment is equal to the drive ring width. A part of the original set flux equal to the drive ring flow now must close on part of the opposite-directed outer flux. Small regions of local closure will be produced in this process and changes will occur in the local closure regions on each side of the drive ring. Again, measurements from Fig. 7 indicate that the net volume of material undergoing flux reversal in this process is of the order of 1.5 times that of the drive ring alone. If there is sufficient drive mmf ring growth will stop only when the width of the drive ring equals the width of the original set ring in Fig. 16(A). In this condition, shown in Fig. 16(C), all the set flux closes outside the region between the two apertures, and an increase in the drive ring can only be obtained if the drive mmf is increased to a value which will cause a growth in the set ring about the larger aperture.

The conclusion that the drive winding must supply not only the energy to create the drive ring but also the energy to create and shift the closure regions is supported by the following experiment. The core was first set. A drive current was applied and increased to a maximum value and then decreased to a low value. The results, shown in Fig. 17, show that a larger drive mmf is required for the initial production of a given output than is required to maintain it. This may be explained as follows: During the first cycle of application of the drive, energy is supplied for the reversal of both the drive ring and the closure regions. On subsequent half-cycles, the drive ring flux may be reversed by the application of a drive which is only sufficient to produce coercive force around the outer periphery of the drive ring.

Now let us consider the effects of simultaneous action of the set and drive mmfs. Suppose that a drive current is established in winding W_2 of an initially blocked core. If a pulse of set mmf F_S is applied when the drive mmf F_D is attempting to reverse the flux in leg 2, the drive and set mmfs will act together to produce a set ring about the larger aperture, as shown in Fig. 16(A). The radius of this set ring will be directly dependent on the peak magnitude of the sum of the

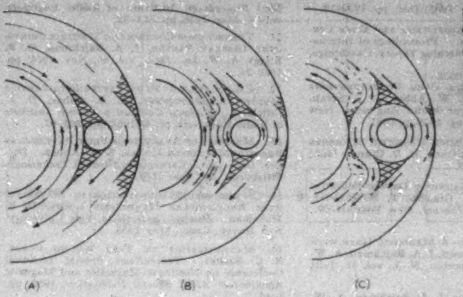


Fig. 16. Flux distribution following application of setting mmf (A), flux distribution after partial drive mmf has been applied (B), and flux distribution for maximum output (C)

set and drive mmfs (F_S+F_D). The production of this set ring requires a hysteresis energy W which is approximately

$$W=B_s\cdot H_c\cdot V \tag{4}$$

where V is the volume of the set ring. The flux ϕ_S of this set ring passes through both the set and drive windings. Thus, the required energy W can be supplied by either or both of these windings since, if W_S and W_D are the energy inputs of the set and drive windings, respectively, and

$$\begin{aligned} W&=W_S+W_D\\ &=\int_o^{\phi_s}F_S d\phi+\int_o^{\phi_s}F_D d\phi\\ &=\int_o^{\phi_s}(F_S+F_D)\,d\phi \end{aligned} \tag{5}$$

In the second half-cycle of the drive, a drive ring of magnetization and closure regions will be produced as shown in Fig. 16(B). The experimental results for the condition just described are those of Figs. 11 and 12.

Next consider the situation for which the experimental results are shown in Fig. 10. The pulse of set mmf is applied when the drive mmf on winding W_2 is attempting to increase the flux in the already saturated leg 2. In a path around the larger aperture periphery the set and drive mmfs are, therefore, in opposition. It might be expected that the setting characteristics would be similar to the previous case except that the output would be dependent on (F_S-F_D). But, from equation 5, it is seen that this would involve a transfer of energy from the set winding to the drive winding. The actual mode of setting the core involves no energy transfer. When a set ring of flux is formed it is simultaneously accompanied by the production of a drive ring if the mmf is greater than the value of 0.85 at necessary to start ring growth. If the drive mmf is sufficiently high the drive ring flux will be equal to the set ring flux as in Fig. 16(C). During this setting process the net change in flux through the drive winding will be zero and the energy input from the drive will also be zero. The energy necessary to produce the set ring, the drive ring, and the accompanying closure regions must all come from the set winding. The minimum set mmf which will start ring growth will, therefore, be about 3.0 +0.85=3.85 AT, a value which corresponds to the zero output in Fig. 10.

Suppose the drive is applied to winding W_3 and the setting pulse during the half-cycle when the drive and setting mmfs are assisting around the core periphery. Under this condition, the flux pattern of Fig. 16(C) will be established during the setting half-cycle. Equal flux changes will occur in each of the windings and the total energy, W, required to establish this pattern will be the sum of the energies from the two windings as in equation 5. The control characteristics for this arrangement are, therefore, similar to those shown in Fig. 11. When the drive and setting mmfs oppose around the core periphery during the setting half-cycle, a set ring as in Fig. 16(A) is produced if $F_S>3.0$ AT. In the next half-cycle, the condition of Fig. 16(B) or 16(C) is obtained. The control characteristics for this arrangement are, therefore, similar to those shown in Fig. 10.

The discussion of the previous four arrangements of the Coincident-Drive-and-Set mode can be extended to cover the situation when the set ring crosses the smaller aperture. It can be shown that there is a transition from an assisting mode of setting to an opposing mode of setting as the setting ring moves from the region of leg 2 to the region of leg 3. This transition accounts for the flat-topped characteristics observed in the curves for all Coincident-Drive-and-Set modes in Figs. 10 and 11.

Conclusion

A number of experiments have been performed on discs made of square-loop ferrite material having one or two circular apertures. The results of these experiments have been used in the development of an approach to the modelling of the various magnetic states which can be created in the disc. The principles used in modelling these flux states are (a) that magnetic flux is established in continuous closed paths within the core, (b) that essentially all parts of the core are at saturation flux density, (c) that the energy required for a change in state is proportional to the volume of material undergoing flux reversal, and (d) that, when there is a choice of state transition, the one involving the minimum energy transfer between windings is chosen.

It is expected that the point of view used in developing these flux-state models will be useful in the analysis of more complex multiaperture devices and will also

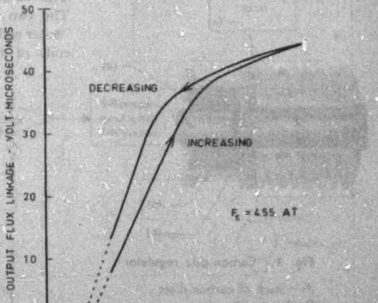


Fig. 17. Output characteristics for increasing, then decreasing drive at given set condition

aid in the development of new multiaperture configurations.

References

1. Magnetics for Computers, A Survey of the Art in 1958, **J. A. Rajchman.** Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication*, 1958, pp. 179–221.

2. Magnetic Materials for Digital Computer Components—I. A Theory of Flux Reversal in Polycrystalline Ferromagnetics, **N. Menyuk, J. B. Goodenough.** *Journal of Applied Physics*, New York, N. Y., vol. 26, Jan. 1955, pp. 8–18.

3. Model for Nonlinear Flux Reversal of Square-Loop Polycrystalline Magnetic Cores, **M. K. Haynes.** *Ibid.*, vol. 29, Mar. 1958, pp. 472–74.

4. Modified Rotational Model of Flux Reversal, **E. M. Gyorgy.** *Ibid.*, Dec., pp. 1709–12.

5. Principles of Transfluxor and Core Circuits, **J. A. Rajchman.** Proceedings of International Symposium on Switching Theory, Cambridge, Mass., Apr. 1956.

6. Multihole Ferrite Core Configurations and Applications, **H. W. Abbot, J. J. Suran.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 45, Aug. 1958, p. 1081.

7. A Logic System Using Magnetic Elements and Connecting Wire Only, **H. D. Crane.** *Ibid.*, vol. 47, Jan. 1959, p. 63.

8. The Laddic—A Magnetic Device for Performing Logic, **U. F. Gianola, T. H. Crowley.** *Bell System Technical Journal*, New York, N. Y., vol. 38, Jan. 1959.

9. The Transfluxor—A Magnetic Gate with Variable Stored Setting, **J. A. Rajchman, A. W. Lo.** *RCA Review*, Princeton, N. J., vol. 16, June 1955, pp. 303–11.

10. The Transfluxor, **J. A. Rajchman, A. W. Lo.** *Proceedings*, Institute of Radio Engineers, vol. 44, Mar. 1956, pp. 321–32.

11. Transfluxor Controlled Electroluminescent Display Panels, **J. A. Rajchman, G. R. Briggs, A. W. Lo.** *Ibid.*, vol. 46, Nov. 1958, pp. 1808–24.

12. Calculation of Flux Patterns in Ferrite Multipath Structures, **S. A. Abbas, D. L. Critchlow.** *National Convention Record*, Institute of Radio Engineers, pt. 4, 1958, pp. 263–67.

13. Methods of Analysis of Flux Patterns in Ferrite Multipath Cores, **S. A. Abbas.** *Doctoral Dissertation*, Carnegie Institute of Technology, Pittsburgh, Pa., June 1958.

14. Superposed Magnetization in Materials with Rectangular Hysteresis Loops, **C. B. Wakeman.** *Doctoral Dissertation*, Yale University, New Haven, Conn., May 1955.

15. Magnetization in Tape Wound Cores, **R. C. Barker.** "Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication*, 1958, pp. 1–46.

The Design of the Generator Voltage Regulator

BERNARD M. VAN EMDEN
ASSOCIATE MEMBER AIEE

FOR APPROXIMATELY 30 years, the aircraft industry has been regulating the output of d-c generators by a carbonpile regulator which employs an electromechanical principle to adjust the field part on d-c generators. The advent of transistors and related solid-state devices, with their attendant high-current capabilities, has made possible the design of a solid-state device equivalent to the carbon pile regulator. This new device offers a number of distinct advantages:

1. Lighter weight.
2. Smaller power dissipation.
3. Greater ruggedness, i.e., capability to withstand environmental conditions such as shock and vibration.
4. Yields equal, or better, regulating characteristics.

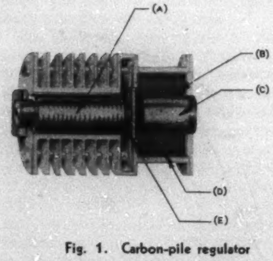


Fig. 1. Carbon-pile regulator

A—Stack of carbon discs
B—Inner coil
C—Iron core
D—Outer coil
E—Diaphragm

The Carbon Pile

The carbon pile presently in use in many aircraft applications is shown in Fig. 1. This figure shows a cutaway view of the carbon-pile regulator, leaving out some of the unnecessary details such as means of adjustment, mounting base, etc. The carbon pile functions as follows: The two coils marked "inner coil" and "outer coil" are the equalizing and control coils of the carbon pile regulator. A current in these coils introduces a magnetomotive force which causes the diaphragm to be depressed against the stack of carbon discs. The pressure on the carbon discs changes their electrical resistance thereby effecting a change in the field current of the generator. The unit is designed so that any change in the generator output voltage correspondingly changes the current in the control coil, the field current, and subsequently effects a like change in the generator output voltage. The spring tension on the diaphragm and the position of the stack of

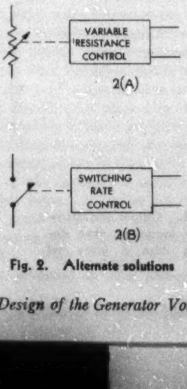


Fig. 2. Alternate solutions

Table I. Performance Data

Serial No.	Output Voltage at No Load	Output Voltage at 120 Amperes	Output Voltage at 200 Amperes
1013	28	27.7	27.5
1014	28	27.7	27.5
1015	28	27.7	27.5
1016	28	27.8	27.5
1017	28	27.7	27.5
1018	28	27.7	27.5
1019	28	27.7	27.5
1020	28	28	27.5
1021	28	27.6	No Test
1022	28	28	27.5
1023	28	27.8	No Test
1024	28	27.6	No Test
1025	28	27.9	No Test
1026	28	27.5	No Test
1027	28	27.5	No Test
1028	28	27.5	27.3
1029	28	27.9	No Test
1030	28	27.8	No Test
1031	28	27.8	No Test
	28 Volts—Direct Current Average	27.731 Average	27.48 Average

Paper 60-1413, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted March 9, 1960; made available for printing December 13, 1960.

Bernard M. Van Emden is with the Automation Development Corporation, Culver City, Calif.

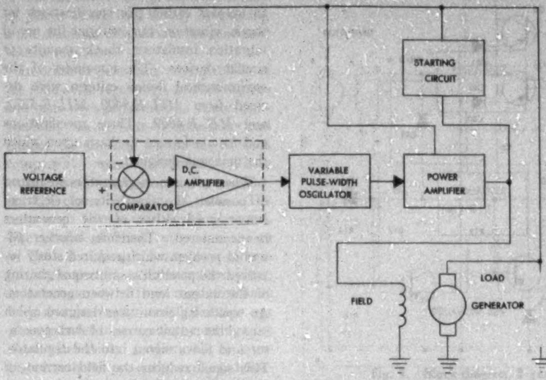


Fig. 3. Block diagram, generator voltage regulator (without equalizer circuit)

carbon discs, as adjusted by the screw on its end, provide a reference for the output of the generator. After some hours of compression and expansion, the carbon discs must be replaced because range of control of the discs has been drastically reduced. Therefore, close inspection and maintenance schedules in high-reliability aircraft are required wherever carbon piles are used.

The other coil on the iron core of the carbon-pile regulator serves as an equalizer coil. Its purpose is to equalize the output voltage when generators are paralleled. It also serves as an additional control device and is connected into the circuit of the paralleled generators in such a manner as to effect the equalization of the output currents of each of the generators. The detail function of equalization will be discussed subsequently.

Solid-State Carbon-Pile Replacement

The solid-state regulator, a replacement for the carbon pile, yields regulation characteristics equivalent, or superior, to the carbon pile as will be shown in this paper. The electromechanical design is such that a carbon-pile socket and existing carbon pile wiring can be utilized to install this regulator. In general, the task of designing a carbon-pile equivalent involved considering and reconsidering the many facets of aircraft power generation. These included:

1. Regulation characteristics.
2. Transient conditions involved in engine starting.
3. Environmental characteristics such as temperature, shock, vibration, etc.
4. Paralleling of generators.
5. Output voltage adjustment.
6. Reliability.

Design Alternatives

Two basic design alternatives presented themselves as shown in Fig. 2. Fig. 2(A) pictures the first of the two design possibilities. A semiconductor device could be built to function as controlled variable resistance. This variable resistance could then be placed in series with the field of the generator, and in this way, generate field current. This would utilize the same basic principle as the carbon pile and, therefore, would involve the same power dissipation problems.

The second alternative is shown in Fig. 2(B). An electronic switch with a controlled switching rate could be placed in series with the field of the d-c generator. The ratio between the "on time" and the "off time" of this switch would then be used as the control device. This ratio (pulse width ratio) could then be varied so that its rate would allow a sufficient average current to flow through the field of the generator in order to yield the proper output voltage. When this type of switching device is used power is dissipated in two basic situations only. The first of these situations occurs when the switch is open. In this case, power is only dissipated by whatever leakage may be involved in this switch. In the second situation, the switch is closed. Under this condition, power is only dissipated by the apparent switch resistance. In addition, if an electronic type of switch is used, power will also be dissipated during the transient switching; i.e., when the switch is going from the "open" or "off" to the "closed" or "on" situation. In the solution of Fig. 2(B), much less power is dissipated than in the variable resistance control of the solution represented in Fig. 2(A). Because of its inherently lower power dissipation, the solution shown in Fig. 2(B) was adopted.

Problem Areas

Before beginning the actual design of the electronic unit, each of the following

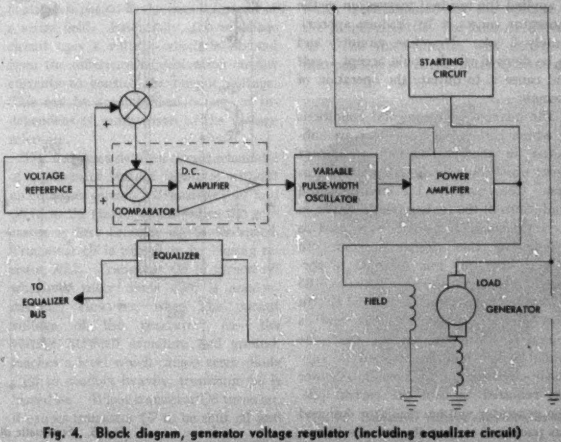


Fig. 4. Block diagram, generator voltage regulator (including equalizer circuit)

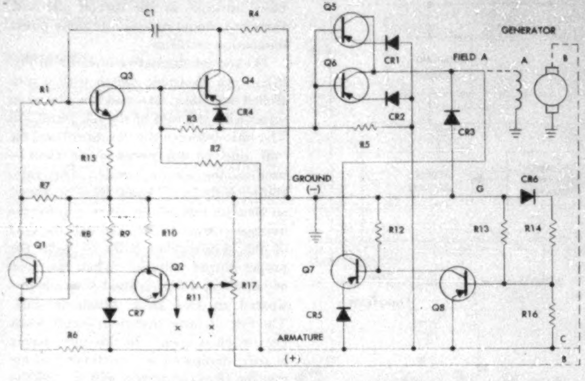


Fig. 5. Schematic diagram, generator voltage regulator

areas had to be fully investigated in detail.

A servomechanism gain analysis aimed at providing the regulation requirement of aircraft military equipment, as well as commercial equipment, was made. The military specification for the carbon pile (*MIL-R-6809*)[1] indicates that a tolerance of ±2 volts, at 28 volts direct current, was the design for the carbon pile. In order to design a substantially improved device a goal of ±1/2 volt, at 28 volts direct current, was established.

The transient conditions, when starting an engine were carefully studied in detail. It was found that when an aircraft of the generator is rotated at proper engine speed, and when no field excitation is applied the residual magnetism in the generator causes it to produce approximately 1 volt. Therefore, circuitry had to be devised which would accept 1 volt and cause it to initiate the operation of the unit.

The adverse environmental conditions to which carbon-pile regulators are subjected in modern aircraft equipment varies widely because of usage and application. Consequently, the environmental limitations of the carbon pile (as given in specification *MIL-R-6809*) were used as criteria in this design program. In general, the unit was designed to perform at a temperature range between −65 F (degrees Fahrenheit) and +160 F. In most aircraft applications the unit is housed in a compartment in which the available air is either trapped or very slowly circulating. As a result, a blower was required to cool the carbon pile. The generator voltage regulator designed was packaged so that the use of a blower was not required except under the most adverse conditions, i.e., a still, trapped air installation. Shock and vibration factors were resolved in compliance with the applicable specification requirements of *MIL-E-5400*[2] (also the design criteria for carbon piles). Through the use of printed circuit techniques and special resistors, an extremely rugged unit was designed which surpasses the carbon-pile shock and vibration specifications. Normally, when the carbon pile is mounted in an aircraft it is placed on a vibration isolation mount. The unit intended to be an electronic equivalent to the carbon pile was designed to surpass carbon-pile specifications for shock, vibration, etc., without the use of vibration insulators, shock mounts or similar devices. The remainder of the environmental design criteria were derived from *MIL-E-5400*, *MIL-E-5272*, and *MIL-R-6809*. These specifications and criteria formed the basis upon which the unit was designed.

When two or more generators are placed in parallel, the possibility of electrical runaway in either of the generators is encountered. Therefore, another potential problem which required study involved the possibility of unequal sharing of the output load between generators. An equalizing circuit was designed which sensed the output current of each generator and fed a signal into the regulator. This signal reduces the field current of whichever generator is carrying the greatest load so that the load current division between generators remains equal.

A potentiometer adjustment for output voltage was selected. Under certain conditions, it may be necessary to use a d-c generator in conjunction with a load which is far removed from the generator. If a carbon-pile regulator were selected for use in this instance, a large diameter wire would have to be procured and installed in order to connect the carbon pile and generator to the load. Not only is this type of setup expensive, but it is often unacceptable because of engineering considerations. The voltage adjustment designed for use in the replacement equipment, however, has an added advantage

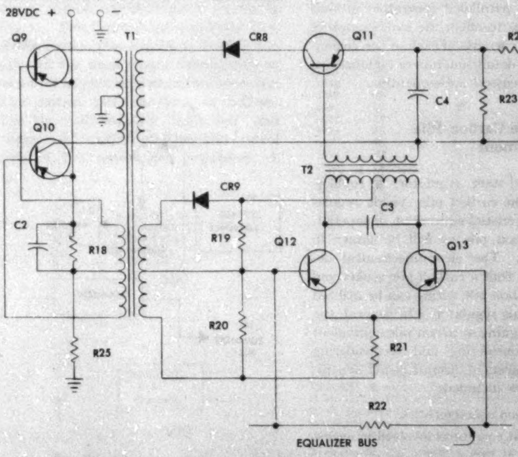


Fig. 6. Schematic diagram, equalizer circuit

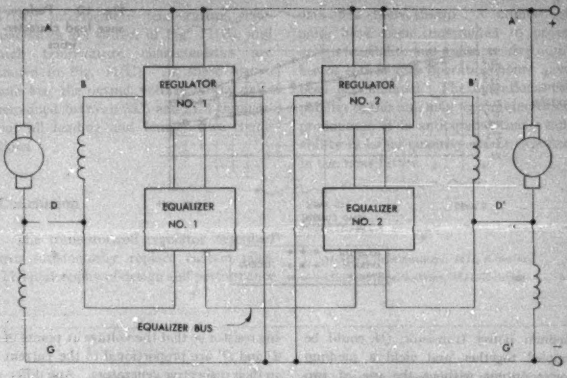


Fig. 7. Block diagram, 2-generator system (without protective relays)

in that the entire voltage adjustment may be placed at the load simply by running a fine diameter wire back to the generator voltage regulator.

In order to determine the reliability of the solid-state regulators, detailed studies and tests were performed on many components. As a result of these tests and studies the limiting parameters of each component were determined. These parameters formed the basis for purchasing the components for 18 units. The units were then constructed and tested. The results of these tests are shown in Table I. A satisfactory rate of uniformity in performance was made evident. A large quantity of these regulators have been and are being, installed within our plant in order to accumulate running time data. Reliability studies will be completed with the presently conducted testing for component failure, and will be followed by possible re-evaluations of component specifications.

Design and Operation

An understanding of the operation of the generator voltage regulator may be gained by the following explanation of its design philosophy and circuitry. Fig. 3 shows a block diagram of the system constructed without an equalizer circuit. As is shown, the output voltage of the generator is compared to a voltage reference in a comparator, and that differential signal is amplified in a d-c amplifier. The output voltage of the d-c amplifier is then used to control the pulse width of a variable pulse-width oscillator. The output of the variable pulse-width oscillator is, in turn, connected to a power amplifier which alternately turns the field current "on" and "off." The resultant pulse-width ratio determines the average field current, which in turn, yields the desired output voltage. A closed loop system, with the generator and load in the loop, is in this way completed. In order to start the operation, a starting circuit was added. The starting circuit connects the armature of the generator to the field of the generator until a desired output voltage is achieved. When this output voltage is obtained, the starting circuit ceases conduction and performs no further functions.

When generators are to be paralleled, a circuit, as indicated by the block diagram of Fig. 4, is utilized. This circuit is the same as the one previously mentioned except that the series field of the generator and the equalizing circuit and comparator were added. When the generators do not contain a series field, an external equalizing resistance conforming to previously mentioned military specifications is placed in the circuit instead of a series field. Essentially, the equalizer circuit uses a voltage which is derived from the difference in generation output currents to control the output voltage. This can be accomplished before, or independent of, comparison to the voltage reference.

Fig. 5 shows a detailed circuit schematic of a generator voltage regulator without an equalizer circuit. Transistors *Q7* and *Q8* form the starter circuit when the generator is first brought up to its speed. Transistor *Q7* is turned on by biasing resistor *R12*. Transistor *Q8* is turned off whenever zener diode *CR6* is nonconducting. However, when the output voltage of the generator, i.e., the voltage between armature and ground, reaches a level which causes zener diode *CR6* to conduct heavily, transistor *Q8* is biased on. When transistor *Q8* turns on, it causes transistor *Q7* to be shut off and a small back bias is developed across transistor *Q7* as a result of rectifier *CR5*. Once zener diode *CR6* conducts, transistor *Q7* is turned off and the starting circuit ceases to draw current. Transistors *Q1* and *Q2* form the d-c amplifier. Zener diode *CR7* serves as the voltage reference for the unit and potentiometer *R17* provides a means for voltage adjustment. The output voltage of the generator is sensed and the difference between this output voltage (divided by the voltage division ratio of *R17*) and the zener voltage of *CR7* is amplified by a d-c amplifier composed of *Q1* and *Q2*. The output voltage of this d-c amplifier is produced, therefore, at the junction of resistors *R1* and *R7*. Transistors *Q3* and *Q4* form a complimentary symmetry, a square-wave multivibrator whose pulse width is determined by the voltage at the junction of *C1*, the base of *Q3* and *R1*. The multivibrator was chosen so that one low power transistor, *Q3*, and one

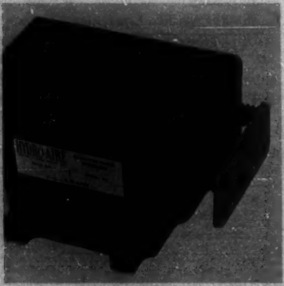


Fig. 8. Generator voltage regulator, overall view

Fig. 9. Generator voltage regulator, opened

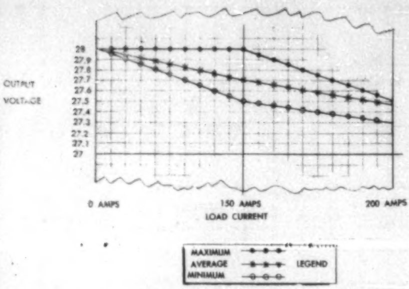


Fig. 10. Performance load characteristics

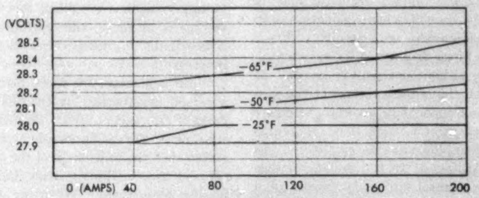
medium power transistor, *Q4*, could be coupled together and yield a medium power output without the use of two medium power transistors. The output of this variable pulse-width oscillator is obtained from the junction of *CR4* and *R3* and is used to drive a power amplifier comprised of *Q5* and *Q6*. This power amplifier acts as a switch between the field and the generator armature. The switching rate is ultimately controlled by the difference between the armature voltage and the zener voltage of *CR7*. The diode which is placed between the field and ground (*CR3*) acts as a "coasting" diode to suppress any inductive transients which might be produced by the collapsing of the generator field when *Q5* and *Q6* are turned off. *CR4*, *CR1*, and *CR2* act to develop a slight back bias across their respective transistors *Q4*, *Q5*, and *Q6*, when these transistors are in the "off," or nonconducting state.

Thus, the system designed employs a d-c amplifier, a variable pulse-width oscillator, and power amplifier techniques which yield an accurate regulation of the output voltages of d-c generators.

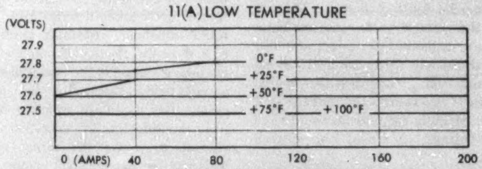
When an equalizer circuit is required, resistor *R11*, shown in Fig. 5, is removed and the circuit shown in Fig. 6 is installed between the points marked *X*. The circuit shown in Fig. 6 is essentially an equalizer circuit. It detects a d-c voltage across resistor *R22* and amplifies this voltage. A reference level is also obtained from this voltage. The amplified d-c voltage is then used as an additional control on the generator output voltage. Transistors *Q9*, *Q10*, and transformer *T1* form a saturable-core square-wave oscillator. The output of this oscillator is used to drive a chopper circuit consisting of *Q12*, *Q13*, and a demodulator circuit (*Q11*). When two or more generators are connected in parallel, wiring is effected in the manner shown in Fig. 7. The series field of each generator acts as an equalizing resistor so that the voltage at points of *D* and *D'* are proportional to the current in their respective generators. Any difference between the voltage at *D* and *D'* causes a current to flow to the equalizer and develop a voltage across *R22* of each equalizer. This current produces a voltage which is fed into the regulator and causes the field current of the generator (which is producing the greatest armature current) to be reduced. Thus, the value of the voltage differential between point *D* and *D'* is reduced. In this way any number of generators can be connected in parallel (see Fig. 7).

The unit, which was constructed according to the described circuit design, is shown in Fig. 8. This photograph shows the power transistors mounted on their appropriate heat sinks and a resistor on the unit. Fig. 9 is a photograph of the opened unit showing a 3-layer package. The first layer consists of the mounting base and can. The second layer is a printed circuit board. The third layer consists of the heat sink and the semiconductor components. An elaborate heat sinking arrangement was necessary in order that the unit can withstand the adverse environmental conditions previously mentioned. The design of the unit is such that the possibility of operational failure from the effects of shock or vibration is remote.

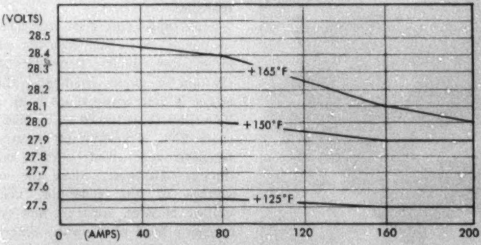
In order to prove the generator voltage



11(A) LOW TEMPERATURE



11(B) NTERMEDIATE TEMPERATURE



11(C) HIGH TEMPERATURE

Fig. 11. Performance characteristics (temperature)

regulator, 19 units were constructed and functionally tested. Fig. 10 shows the performance load characteristics of these units. The curve marked with the asterisks indicates the average performance of all the units tested up to this time. The curve marked with the solid dot indicates the upper limit of the units tested and the curve marked with the open dots indicates the lower limit of the tests.

Table I gives the information from which Fig. 10 was derived. Fig. 11 indicates a typical temperature variation for one of the regulator units. Low temperature characteristics are shown in Fig. 11(A), intermediate temperature characteristics are shown in Fig. 11(B), and high temperature characteristics are shown in Fig. 11(C). As these figures indicate, the output voltage of the units remained between 28.5 and 27.5 throughout all loading and temperature conditions.

Conclusion

The transistorized regulator described can satisfactorily replace carbon piles. The philosophy of design and performance data has been shown. A number of units have been constructed to prove design feasibility and an impressive number of satisfactory operating hours have been accumulated. The operational capability of this unit may warrant increased production, it is anticipated that availability in larger quantity can be expected in the near future.

References

1. Military Specifications, *MIL-R-Series.*
2. Military Specifications, *MIL-E-Series.*

A 4-Megacycle 24-Bit Checked Binary Adder

M. E. HOMAN
NONMEMBER AIEE

RECENTLY DEVELOPED large computers use multiple memory units and buffering between memory and processing to achieve a computation rate higher than the memory rate. In such cases computer speed is limited by the basic adders employed. Because these adders represent a small fraction of the total machine cost, it is economical to use the fastest parallel design available. Checking, if provided, should be accomplished in such a way that the add-cycle time is not increased. This implies parallel rather than serial operation of adder and checking logic. In this paper, the logical design of a 24-bit binary adder developed for the IBM (International Business Machines Corporation) Stretch computer project is described. Direct-coupled type logic is used and high performance is achieved by means of a parallel design employing fast switching devices. Checking employs a combination of parity and duplicate-carry logic for single-error detection.

Solid-state, nsec (nanosecond) logical elements employing diffused-junction transistors are used in current-mode and emitter-follower configurations. The cascading factors, level-setting characteristics, and relative speeds of these circuits are taken into account in the logical design. A detailed analysis of the logical design is carried out subsequently in terms of 4-bit groupings tied together by a parallel carry-look-ahead system (second-level carry logic). Bit functions, internal and external group functions, carry look-ahead, sum generation, and checking are covered in some detail. Only the degree of logical duplication and the use of single-input emitter followers to achieve the required cascading factors is omitted for the sake of clarity. Latching, storage, and gating are included in a general way.

Storage, gating, and half-sum generation elements are parity checked. Six-bit parity groups are used to keep parity trees short and to facilitate fault location. Carry and sum generation are checked by redundant serial carry logic operating in an "after the fact" mode. Checking logic is performed in parallel with adding and storage operations and therefore does not penalize performance. In effect, the checking cycle lags the data cycle by approximately 1/4 cycle. Carry errors are detected on a single-bit basis. Additions are completed in a relatively fixed time interval because of the parallel-carry system and the nature of the logical elements employed. Data timing provides no special advantage where such characteristics hold and a clock source of symmetrical samples is provided for control. As will be shown, a complete add operation, including storage in transistor registers, can be performed in 250 nsec.

Only 4-bit groupings are described in the detailed analysis. As is apparent from the Boolean expressions, the optimum group size is primarily a function of cascading factors and maximum logical-block size. In the actual design covered briefly at the end of this paper, 2-bit, 3-bit, and 4-bit groups are used and two basic add modes are provided. One mode provides for separate simultaneous operation of 6-bit and 18-bit sections. The second mode provides for 24-bit operation. Other control inputs provide a "plus one" for counting up or for subtracting and a "minus one" function for counting down. The adder path may also be used to check register-to-register transfers of data. A fast overflow carry output from the low-order 6-bit section is provided as a supplementary output.

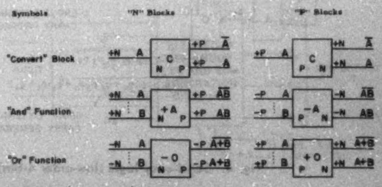


Fig. 1 Characteristics of current mode circuit

Paper 61-193, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960, and re-presented for discussion only at the Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 11, 1960; made available for printing December 13, 1960.

M. E. Homan is with the International Business Machines Corporation, Poughkeepsie, N. Y.

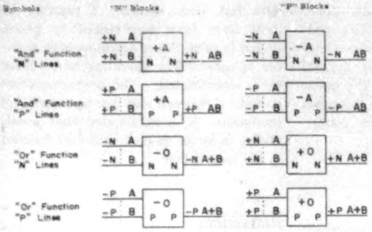
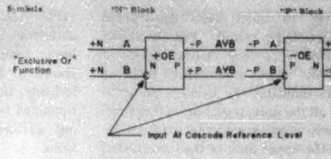

Symbols
"N" Blocks
"P" Blocks
"And" Function "N" Lines
+N A
+N B
+A
N N
+N AB
−N A
−N B
−A
N N
−N AB
"And" Function "P" Lines
+P A
+P B
+A
P P
+P AB
−P A
−P B
−A
P P
−P AB
"Or" Function "N" Lines
−N A
−N B
−O
N N
−N A+B
+N A
+N B
+O
N N
+N A+B
"Or" Function "P" Lines
−P A
−P B
−O
P P
−P A+B
+P A
+P B
+O
P P
+P A+B

Fig. 2 (left). Characteristics of emitter-follower circuit

Fig. 3 (right). Characteristics of cascode Exclusive-Or circuit


Symbols
"N" Block
"P" Block
"Exclusive Or" Function
+N A
+N B
+OE
N P
−P A⊽B
+P A⊽B
−P A
−P B
−OE
P N
+N A⊽B
−N A⊽B
Input At Cascode Reference Level

Circuit Characteristics and Symbolism

I. Current-Mode Circuits

(See Fig. 1)

1. One to eight inputs per logic block.
2. Both inphase and out-of-phase outputs.
3. Output levels reset (gain = 1).
4. Average logical delay ≈20 nsec per stage.
5. Output can drive up to three other circuits.
6. *N*-type blocks use p-n-p transistors to perform +AND or −OR functions.
7. *P*-type blocks use n-p-n transistors to perform +OR or −AND functions.
8. *N* blocks can drive only *P* blocks or emitter-follower circuits.
9. *P* blocks can drive only *N* blocks or emitter-follower circuits.

(Note that in Fig. 1 both positive and negative logical symbols are used and that lines are designated *N* or *P* in accordance with the type of block they can drive. The *N* lines are referenced to ground, whereas the *P* lines are referenced to −6 volts.)

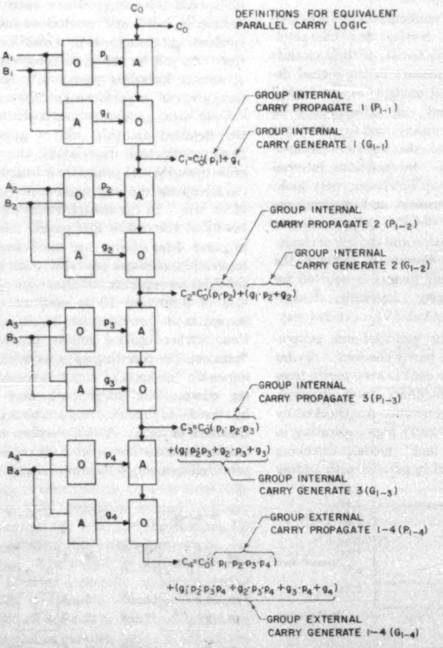

C0
C0
DEFINITIONS FOR EQUIVALENT PARALLEL CARRY LOGIC
A1
B1
O
p1
A
A
g1
O
GROUP INTERNAL CARRY PROPAGATE 1 (P1−1)
GROUP INTERNAL CARRY GENERATE 1 (G1−1)
C1=C'0(p1)+g1
A2
B2
p2
g2
GROUP INTERNAL CARRY PROPAGATE 2 (P1−2)
GROUP INTERNAL CARRY GENERATE 2 (G1−2)
C2=C'0(p1·p2)+(g1·p2+g2)
A3
B3
p3
g3
GROUP INTERNAL CARRY PROPAGATE 3 (P1−3)
C3=C'0(p1·p2·p3)
+(g1·p2p3+g2·p3+g3)
GROUP INTERNAL CARRY GENERATE 3 (G1−3)
A4
B4
p4
g4
GROUP EXTERNAL CARRY PROPAGATE 1−4 (P1−4)
C4=C'0(p1·p2·p3·p4)
+(g1·p2·p3·p4+g2·p3·p4+g3·p4+g4)
GROUP EXTERNAL CARRY GENERATE 1−4 (G1−4)

Fig. 4. Serial-carry logic (low-order 4-bits)

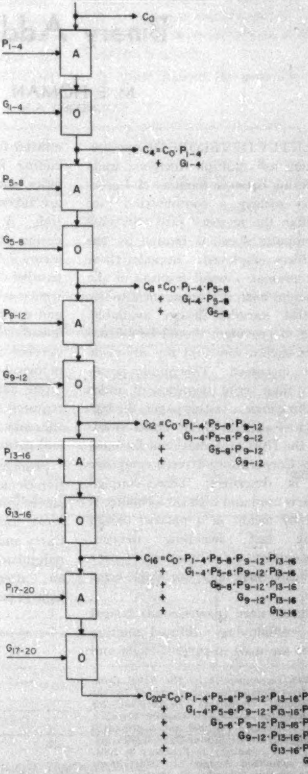

C0
C0
P1−4
A
G1−4
O
C4=C0·P1−4
+ G1−4
P5−8
G5−8
C8=C0·P1−4·P5−8
+ G1−4·P5−8
+ G5−8
P9−12
G9−12
C12=C0·P1−4·P5−8·P9−12
+ G1−4·P5−8·P9−12
+ G5−8·P9−12
+ G9−12
P13−16
G13−16
C16=C0·P1−4·P5−8·P9−12·P13−16
+ G1−4·P5−8·P9−12·P13−16
+ G5−8·P9−12·P13−16
+ G9−12·P13−16
+ G13−16
P17−20
G17−20
C20=C0·P1−4·P5−8·P9−12·P13−16·P17−20
+ G1−4·P5−8·P9−12·P13−16·P17−20
+ G5−8·P9−12·P13−16·P17−20
+ G9−12·P13−16·P17−20
+ G13−16·P17−20
+ G17−20

Fig. 5. Serial-carry logic for 4-bit groups

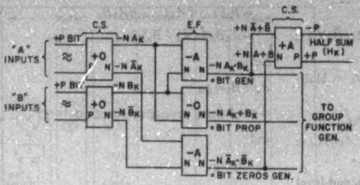

C.S.
E.F.
C.S.
"A" INPUTS
"B" INPUTS
+P BIT
+P BIT
+0
P N
−N A_K
−N Ā_K
−N B_K
−N B̄_K
−A
N N
−O
N N
−A
N N
+N Ā+B̄
+N A+B
+A
N P
−P
HALF SUM
+P (H_K)
−N A_K·B_K
• BIT GEN
−N A_K+B_K
• BIT PROP
−N Ā_K·B̄_K
• BIT ZEROS GEN.
TO GROUP FUNCTION GEN.

Fig. 6. (left). Bit function generation (per bit)

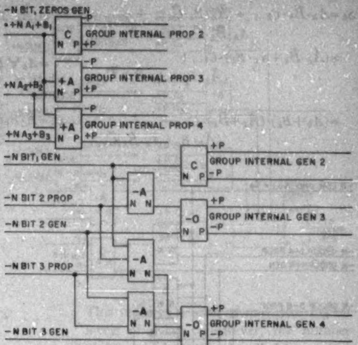

−N BIT, ZEROS GEN
+N A₁+B₁
+N A₂+B₂
+N A₃+B₃
C
+A
+A
N P
−P
+P
GROUP INTERNAL PROP 2
GROUP INTERNAL PROP 3
GROUP INTERNAL PROP 4
−N BIT₁ GEN
−N BIT 2 PROP
−N BIT 2 GEN
−N BIT 3 PROP
−N BIT 3 GEN
−A
N N
−O
N P
GROUP INTERNAL GEN 2
GROUP INTERNAL GEN 3
GROUP INTERNAL GEN 4

Fig. 7 (right). Group internal function generation (per 4-bit group)

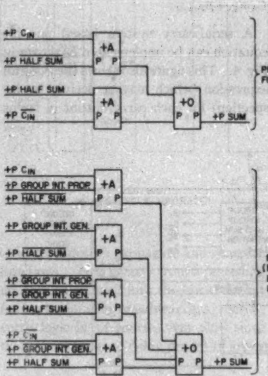

+P C_IN
+P HALF SUM
+P HALF SUM
+P C̄_IN
+A
P P
+O
P P
+P SUM
PER LOW ORDER BIT FOR EACH GROUP
+P C_IN
+P GROUP INT. PROP
+P HALF SUM
+P GROUP INT. GEN.
+P HALF SUM
+P GROUP INT. PROP
+P GROUP INT. GEN.
+P HALF SUM
+P C̄_IN
+P GROUP INT. GEN.
+P HALF SUM
+A
P P
+O
P P
+P SUM
PER BIT (EXCEPT FOR LOW ORDER POS. IN EACH GROUP)

Fig. 8 (left). Sum generation

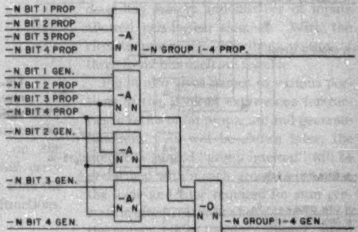

−N BIT 1 PROP
−N BIT 2 PROP
−N BIT 3 PROP
−N BIT 4 PROP
−A
N N
−N GROUP 1−4 PROP.
−N BIT 1 GEN.
−N BIT 2 PROP
−N BIT 3 PROP
−N BIT 4 PROP
−N BIT 2 GEN.
−N BIT 3 GEN.
−N BIT 4 GEN.
−O
N N
−N GROUP 1−4 GEN.

Fig 9. (right). Group external function generation (per 4-bit group)

II. EMITTER-FOLLOWER CIRCUITS

(See Fig. 2)

1. One to six inputs per logic block.
2. Inphase output only.
3. Maximum chain length of four (gain <1).
4. Average logical delay ≈10 nsec per stage.
5. Output can drive up to three other circuits.
6. Current-switching blocks are used for level setting.
7. Nontranslating. Output line type same as inputs.

Single-input blocks may be used as fast 3-base drivers.

III. CASCODE EXCLUSIVE-OR CIRCUITS

(See Fig. 3)

1. Only two inputs required, one at normal current-switch reference level and one at a special cascode reference level.
2. Current-switching output must be used to drive cascode input.
3. Up to three cascode inputs can be driven from a logic-block output.
4. Both inphase and out-of-phase outputs. (Equivalent to current-switching outputs.)
5. Output levels reset (gain = 1).
6. Average logical delay ≈30 nsec per stage.
7. Output can drive up to three other circuits.
8. *N*-type blocks use p-n-p transistors to perform +EXCLUSIVE OR function.
9. *P*-type blocks use n-p-n transistors to perform −EXCLUSIVE OR function.
10. *N* blocks can drive only *P* blocks or emitter-follower circuits.
11. *P* blocks can drive only *N* blocks or emitter-follower circuits.

Analysis of Adder Design

An analysis similar to that of Weinberger and Smith[1] leading to the design of a parallel adder will now be described in detail.

Let:

A = augend

$= A_k 2^{k-1} + A_{k-1} 2^{k-2} + \ldots + A_1$ 2 degrees

B = addend

$= B_k 2^{k-1} + B_{k-1} 2^{k-2} + \ldots + B_1$ 2 degrees

S = sum

$= S_k 2^{k-1} + S_{k-1} 2^{k-2} + \ldots + S_1$ 2 degrees

and C_k = the carry resulting from the addition in the *k*th digit position. The rules for binary addition are:

Function Table for Binary Addition

augend, A_k	0	0	0	0	1	1	1	1
addend, B_k	0	0	1	1	0	0	1	1
Previous Carry, C_{k-1}	0	1	0	1	0	1	0	1
sum, S_k	0	1	1	0	1	0	0	1
carry, C_k	0	0	0	1	0	1	1	1

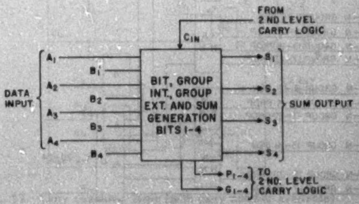

FROM 2 ND LEVEL CARRY LOGIC
C_IN
DATA INPUT
A₁
B₁
A₂
B₂
A₃
B₃
A₄
B₄
BIT, GROUP INT., GROUP EXT. AND SUM GENERATION BITS 1-4
S₁
S₂
S₃
S₄
SUM OUTPUT
P₁₋₄
G₁₋₄
TO 2 ND. LEVEL CARRY LOGIC

Fig. 10. Block representation of 4-bit group

$$S_k = \bar{A}_k \cdot \bar{B}_k \cdot C_{k-1} + \bar{A}_k \cdot B_k \cdot \bar{C}_{k-1} + A_k \cdot \bar{B}_k \cdot \bar{C}_{k-1} + A_k \cdot B_k \cdot C_{k-1}$$

$$= (\bar{A}_k \cdot \bar{B}_k + A_k \cdot B_k) \cdot C_{k-1} + (\bar{A}_k \cdot B_k + A_k \cdot \bar{B}_k) \cdot \bar{C}_{k-1}$$

$$= \overline{(A_k + B_k) \cdot (\bar{A}_k + \bar{B}_k)} \cdot C_{k-1} + (\bar{A}_k \cdot B_k + A_k \cdot \bar{B}_k)\bar{C}_{k-1}$$

$$= \overline{(\bar{A}_k \cdot B_k + A_k \cdot \bar{B}_k)} \cdot C_{k-1} + (\bar{A}_k \cdot B_k + A_k \cdot \bar{B}_k)\bar{C}_{k-1}$$

$$= A_k \forall B_k \forall C_{k-1} = H_k \forall C_{k-1}$$

where $\forall$ = EXCLUSIVE-OR and H_k = Half-sum_k

$$C_k = \bar{A}_k B_k C_{k-1} + A_k \bar{B}_k C_{k-1} + A_k B_k \bar{C}_{k-1} + A_k B_k C_{k-1}$$

$$= B_k C_{k-1} + A_k C_{k-1} + A_k B_k$$

$$= A_k B_k + (A_k + B_k) C_{k-1}$$

Let:

$$A_k + B_k = p_k$$

= Bit *k* carry propagate function

and:

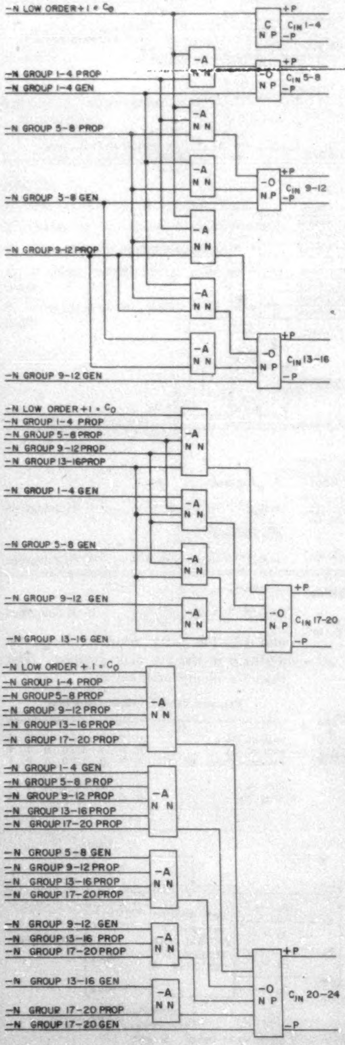
$$A_k \cdot B_k = g_k$$

= Bit *k* carry generate function

Then

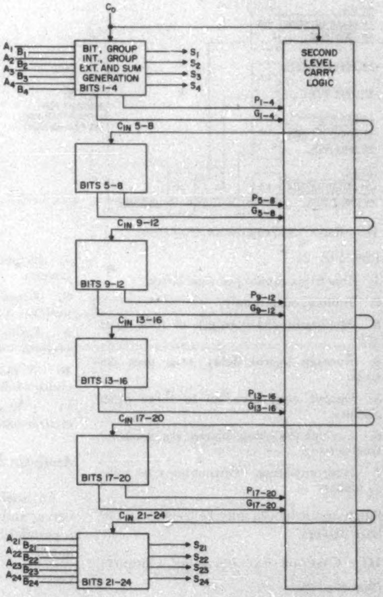
$$C_k = g_k + p_k C_{k-1}$$

A serial-carry system based on this equation can be implemented as shown in Fig. 4. This figure also shows the Boolean expression (which may be derived by inspection) for each carry output in terms



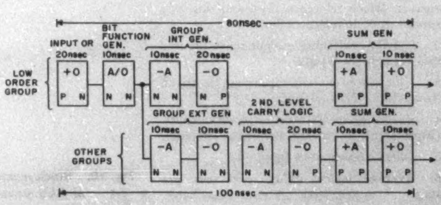
-N LOW ORDER + 1 = C0
-N GROUP 1-4 PROP
-N GROUP 1-4 GEN
-N GROUP 5-8 PROP
-N GROUP 5-8 GEN
-N GROUP 9-12 PROP
-N GROUP 9-12 GEN
-N GROUP 13-16 PROP
-N GROUP 13-16 GEN
-N GROUP 17-20 PROP
-N GROUP 17-20 GEN
C
N P
-A
N N
-O
N P
+P
-P
CIN 1-4
CIN 5-8
CIN 9-12
CIN 13-16
CIN 17-20
CIN 20-24

Fig. 11 (left). Second-level carry logic



C0
A1 B1
A2 B2
A3 B3
A4 B4
BIT, GROUP INT., GROUP EXT. AND SUM GENERATION BITS 1-4
S1
S2
S3
S4
SECOND LEVEL CARRY LOGIC
P1-4
G1-4
CIN 5-8
BITS 5-8
P5-8
G5-8
CIN 9-12
BITS 9-12
P9-12
G9-12
CIN 13-16
BITS 13-16
P13-16
G13-16
CIN 17-20
BITS 17-20
P17-20
G17-20
CIN 21-24
A21 B21
A22 B22
A23 B23
A24 B24
BITS 21-24
S21
S22
S23
S24

Fig. 12 (right). Block representation of 24-bit adder using 4-bit groups



80nsec
INPUT OR 20nsec
BIT FUNCTION GEN. 10nsec
GROUP INT GEN.
10nsec
20nsec
SUM GEN
10nsec
10 nsec
LOW ORDER GROUP
+O
P N
A/O
N N
-A
N N
-O
N P
+A
P P
+O
P P
GROUP EXT GEN
10nsec
10nsec
2ND LEVEL CARRY LOGIC
10nsec
20 nsec
SUM GEN.
10nsec
10nsec
OTHER GROUPS
100nsec

Fig. 13 (below right). Path length and logical delay through adder

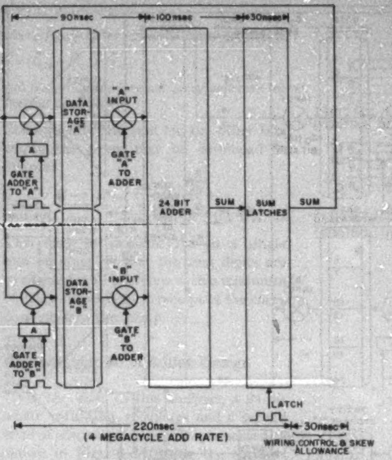

90nsec
100nsec
30nsec
DATA STORAGE "A"
"A" INPUT
GATE "A" TO ADDER
A
GATE ADDER TO "A"
24 BIT ADDER
SUM
SUM LATCHES
SUM
DATA STORAGE "B"
"B" INPUT
GATE "B" TO ADDER
A
GATE ADDER TO "B"
LATCH
220nsec
(4 MEGACYCLE ADD RATE)
30nsec
WIRING CONTROL & SKEW ALLOWANCE

Fig. 14 (left). Block diagram of adder with sum latches and storage

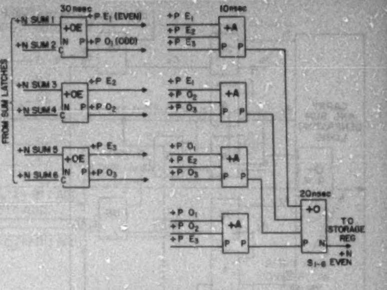

FROM SUM LATCHES
30nsec
+N SUM 1
+N SUM 2
+OE
+P E1 (EVEN)
+P O1 (ODD)
+N SUM 3
+N SUM 4
+P E2
+P O2
+N SUM 5
+N SUM 6
+P E3
+P O3
10nsec
+A
20nsec
+O
TO STORAGE REG
+N S1-6 EVEN

Fig. 15 (right). Six-bit parity generation

of lower order bit propagate and generate functions. C_0 is a carry from a preceding bit or a "plus one" into the low-order bit for use in counting or subtracting. Serial-carry logic is, of course, very slow since two levels of logical delay per bit of adder must be allowed for a possible carry propagation from C_0 to the high-order bit. Parallel implementation, on the other hand, can produce the full carry number in a single AND-OR stage from the bit propagate and generate functions. This requires a total of only two logical levels of delay. However, the number of bits which can be handled by such a simple arrangement is limited by the cascading factors and number of inputs allowed per logical element. With the circuits previously described, groups of three to six bits each are feasible.

Fig. 4 also gives names to various portions of the Boolean expressions for carries in terms of bit propagate and generate functions. As will be shown later, the functions labeled "group internal" will be preformed within each group to minimize the logic and time required for sum generation after a carry has been generated for the groups. The "group external" functions provide the input information for a second level of carry logic that treats groups of bits just as we have treated individual bits in the carry logic above.

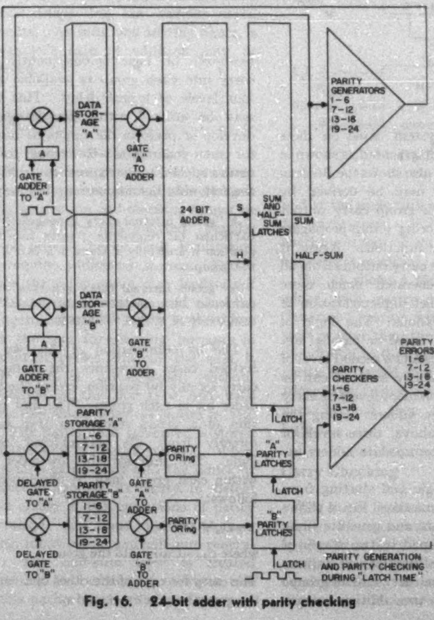

DATA STORAGE "A"
GATE "A" TO ADDER
A
GATE ADDER TO "A"
24 BIT ADDER
S
H
SUM AND HALF-SUM LATCHES
SUM
HALF-SUM
PARITY GENERATORS 1-6 7-12 13-18 19-24
DATA STORAGE "B"
GATE "B" TO ADDER
GATE ADDER TO "B"
PARITY CHECKERS 1-6 7-12 13-18 19-24
PARITY ERRORS 1-6 7-12 13-18 19-24
LATCH
PARITY STORAGE "A"
1-6
7-12
13-18
19-24
DELAYED GATE TO "A"
GATE "A" TO ADDER
PARITY ORing
"A" PARITY LATCHES
LATCH
PARITY STORAGE "B"
DELAYED GATE TO "B"
GATE "B" TO ADDER
PARITY ORing
"B" PARITY LATCHES
LATCH
PARITY GENERATION AND PARITY CHECKING DURING "LATCH TIME"

Fig. 16. 24-bit adder with parity checking

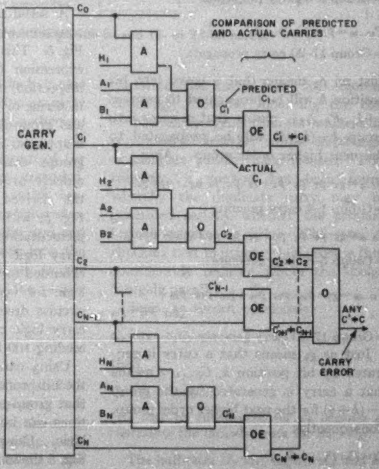

C0
COMPARISON OF PREDICTED AND ACTUAL CARRIES
H1
A1
B1
A
O
PREDICTED C1
C1'
CARRY GEN.
C1
OE
C1' ≠ C1
ACTUAL C1
H2
A2
B2
C2'
C2
C2' ≠ C2
CN-1
C'N-1
C'N-1 ≠ CN-1
O
ANY C' ≠ C
CARRY ERROR
HN
AN
BN
CN'
CN
CN' ≠ CN

Fig. 17 Carry checking logic (with carry number available)

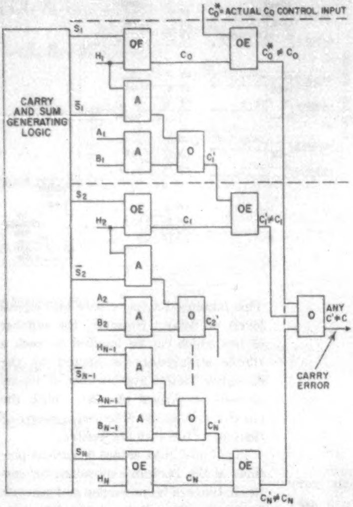

C_0^* = ACTUAL C_0 CONTROL INPUT
CARRY AND SUM GENERATING LOGIC
S_1
H_1
OE
C_0
$C_0^* \neq C_0$
A
$\bar{S}_1$
A_1
B_1
O
C_1'
S_2
H_2
C_1
$C_1' \neq C_1$
$\bar{S}_2$
A_2
B_2
C_2'
H_{N-1}
$\bar{S}_{N-1}$
A_{N-1}
B_{N-1}
C_N'
S_N
H_N
C_N
$C_N' \neq C_N$
ANY $C' \neq C$
CARRY ERROR

Fig. 18. Carry and sum checking logic (carry number not available)

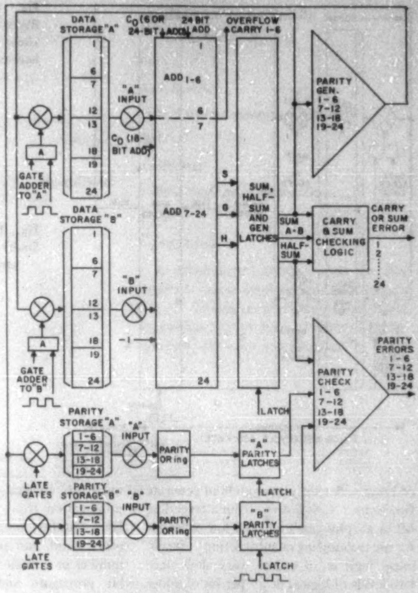

DATA STORAGE "A"
C_0 (6 OR 24-BIT ADD)
24 BIT ADD
OVERFLOW CARRY 1-6
1
6
7
12
13
18
19
24
"A" INPUT
ADD 1-6
C_0 (18-BIT ADD)
ADD 7-24
A
GATE ADDER TO "A"
DATA STORAGE "B"
"B" INPUT
-1
GATE ADDER TO "B"
S
G
H
SUM, HALF-SUM AND GEN LATCHES
PARITY GEN. 1-6 7-12 13-18 19-24
SUM A·B
HALF-SUM
CARRY & SUM CHECKING LOGIC
CARRY OR SUM ERROR
1
2
24
PARITY CHECK 1-6 7-12 13-18 19-24
PARITY ERRORS 1-6 7-12 13-18 19-24
LATCH
PARITY STORAGE "A"
1-6
7-12
13-18
19-24
"A" INPUT
PARITY ORing
"A" PARITY LATCHES
LATE GATES
PARITY STORAGE "B"
"B" INPUT
"B" PARITY LATCHES

Fig. 19. 24-bit checked binary adder

To divide the adder into 4-bit groups, one of which is shown in Fig. 4, let

$P_{1-4} = p_1 \cdot p_2 \cdot p_3 \cdot p_4$

= Group 1–4 carry propagate

$P_{5-8} = p_5 \cdot p_6 \cdot p_7 \cdot p_8$

= Group 5–8 carry propagate

⋮

$P_{17-20} = p_{17} \cdot p_{18} \cdot p_{19} \cdot p_{20}$

= Group 17–20 carry propagate

Just as p_k means that a carry into bit position k will be propagated to position $k+1$, $P_{k-(k+4)}$ means that a carry into group $k-(k+4)$ will be propagated to the next higher order group. Also:

$G_{1-4} = g_1 \cdot p_2 \cdot p_3 \cdot p_4 + g_2 \cdot p_3 \cdot p_4 + g_3 \cdot p_4 + g_4$

= Group 1–4 carry generate

$G_{5-8} = g_5 \cdot p_6 \cdot p_7 \cdot p_8 + g_6 \cdot p_7 \cdot p_8 + g_7 \cdot p_8 + g_8$

= Group 5–8 carry generate

⋮

$G_{17-20} = g_{17} \cdot p_{18} \cdot p_{19} \cdot p_{20} + g_{18} \cdot p_{19} \cdot p_{20} + g_{19} \cdot p_{20} + g_{20}$

= Group 17–20 carry generate

Just as g_k means that a carry is generated by bit position k, $G_{k-(k+4)}$ means that a carry is generated by the group $k-(k+4)$ for the next higher order group. Consequently:

$C_4 = C_0 \cdot P_{1-4} + G_{1-4}$

$C_8 = C_4 \cdot P_{5-8} + G_{5-8}$

$C_{12} = C_8 \cdot P_{9-12} + G_{9-12}$

$C_{16} = C_{12} \cdot P_{13-16} + G_{13-16}$

$C_{20} = C_{16} \cdot P_{17-20} + G_{17-20}$

A serial-carry system based on these equations can be implemented as shown in Fig. 5. This figure also shows the Boolean expression (which may be derived by inspection) for each group carry output in terms of lower order group-propagate and group-generate functions. Again, it is apparent that the carry outputs from all groups can be generated much more quickly by a parallel implementation of the derived expressions. The logic of Fig. 5, with either serial or parallel implementation, represents a second level of carry logic. The above analysis can be extended to three or more levels of carry logic for very large adders. Using the circuits described above, three levels of carry logic can accommodate adders exceeding 100 bits.

Using parallel logic and starting from the bit-propagate functions, Fig. 4 shows that group-propagate and generate functions can be developed in two stages of logic. Using these functions as inputs, Fig. 5 shows that the carry into all groups can be developed in two additional stages (AND-OR) of logic. Consequently the carry into each group is available after four levels of logical delay. This time may be utilized within the groups to develop or preform the functions needed for sum generation. Returning to the names applied to the expressions in Fig. 4, the following definitions may be derived:

A "group internal carry propagate" to a particular bit represents a carry into that position if and only if there is a carry into the group.

A "group internal carry generate" to a particular bit represents a carry into that position from within the group.

The OR combination of the two described cases represents the complete carry for any bit position, except the low-order bit in each group. The first position of a group receives a true carry directly from the second-level carry logic. The sum for the low-order bit within each group may be developed as follows:

$S = C_{in} \forall H$

where C_{in} = Carry into the group

The carry for each of the other bits within the group may be expressed as:

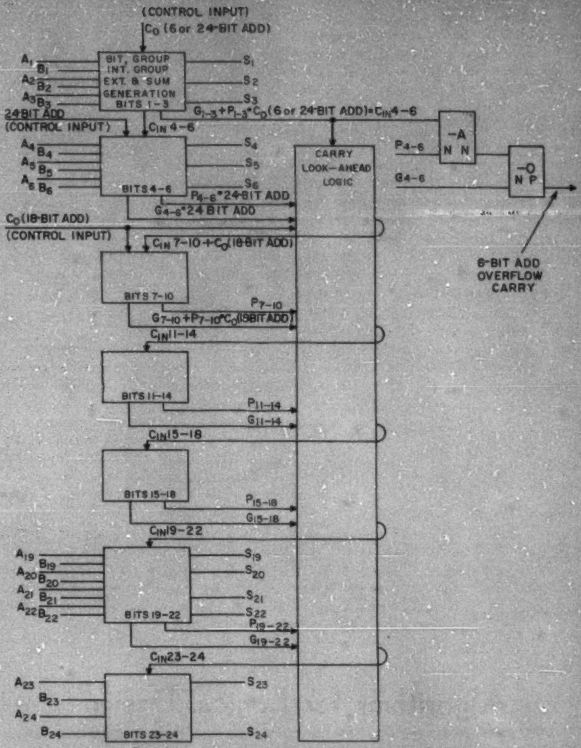

(CONTROL INPUT)
C₀ (6 or 24-BIT ADD)
BIT, GROUP INT. GROUP EXT. & SUM GENERATION BITS 1-3
A1
B1
A2
B2
A3
B3
S1
S2
S3
G1-3+P1-3·C0(6 or 24-BIT ADD)=CIN4-6
24-BIT ADD (CONTROL INPUT)
CIN 4-6
A4
B4
A5
B5
A6
B6
BITS 4-6
S4
S5
S6
P4-6·24-BIT ADD
G4-6·24 BIT ADD
CARRY LOOK-AHEAD LOGIC
P4-6
G4-6
-A N N
-O N P
6-BIT ADD OVERFLOW CARRY
C0 (18-BIT ADD) (CONTROL INPUT)
CIN 7-10+C0(18-BIT ADD)
BITS 7-10
P7-10
G7-10+P7-10·C0(18BIT ADD)
CIN11-14
BITS 11-14
P11-14
G11-14
CIN15-18
BITS 15-18
P15-18
G15-18
CIN19-22
A19
B19
A20
B20
A21
B21
A22
B22
BITS 19-22
S19
S20
S21
S22
P19-22
G19-22
CIN23-24
A23
B23
A24
B24
BITS 23-24
S23
S24

Fig. 20. Adder design with 6- and 18- or 24-bit modes

$$C = C_{in} \cdot P_{int} + G_{int}$$

where P_{int} = Group internal carry propagate

$$\bar{C} = (\bar{C}_{in} + \bar{P}_{int}) \cdot \bar{G}_{int}$$

and G_{int} = Group internal carry generate for that bit position

Consequently the sum for the other bits within the group may be developed as follows:

$$S = C \forall H = C_{in} P_{int} \cdot \bar{H} + G_{int} \cdot \bar{H} + \bar{P}_{int} \cdot \bar{G}_{int} \cdot H + \bar{C}_{in} \cdot \bar{G}_{int} \cdot H$$

This may be implemented in a single AND-OR stage so that the sum digits are available for all positions with a minimum of delay subsequent to receipt of the carry decision into the group.

Implementation of Adder Design

On the basis of this analysis, a 24-bit adder with 4-bit groupings and a second level of carry logic may be implemented as shown in Figs. 6–14, using the current-switching and emitter-follower circuits already described. Although not shown, heavy loading conditions are met by using 1-input emitter-follower circuits as 3-base power drivers where additional delay can be tolerated, and by duplicating circuitry in other cases. Fig. 14 shows a block diagram of the adder, including logical delays for the various serial elements. As indicated in this figure, a 4-megacycle rate of addition may be achieved.

Parity Checking

For parity checking, the adder is divided into four groups of six bits each. Larger groups are not suitable because parity generation and checking cannot be completed in the time allotted. Smaller groups provide better fault localization but require additional storage. Fortunately the 6-bit group size lends itself to the parity generation and checking arrangements used. Parity is generated from the sum for checking purposes in subsequent cycles wherein the sum is used. This is accomplished without translation of the line type and with reasonably small delay, as shown in Fig. 15. The parity thus generated is stored with the sum in trigger storage. When the sum is gated back into the adder, it is checked against its parity to detect single errors in the operations of parity generation, sum and parity storage, gating, input ORing, and half-sum generation. The half-sum logic is treated as the first level of an EXCLUSIVE-OR tree for this parity check. This reduces cost and at the same time checks the half-sum for single errors. In the section on carry checking it will be shown that a checked half-sum is essential.

Fig. 16 shows a block diagram of the adder with parity checking included. Generated parity bits are gated into storage later than the sum and therefore do not add to the cycle. If the sum and parity are gated directly back into the adder, the parity bits catch up to the data because they bypass the adder logic.

Carry Checking

The use of limited redundancy, such as parity, for economical checking is not applicable to carry logic because the carry number has no simple relationship, such as odd/even, to the input numbers. For this reason it is usually necessary to duplicate the carry logic for rigorous checking. However, doubling the cost of the carry generation is unpleasant regardless of whether simple serial logic or complex parallel logic is used. Fortunately, a method has been found whereby the duplicate carry may be generated serially whether the primary logic is parallel or serial. The ratio of checking cost to primary logic cost is thus considerably reduced where the design is highly parallel.

Since, as shown previously,

$$C_k = g_k + p_k C_{k-1}$$

then

$$C_k = g_k + H_k C_{k-1}$$

(H_k may be substituted for p_k because g_k represents the inclusive case for $A_k + B_k$.)

The half-sum H_k and inputs A_k and B_k have been parity checked as shown above and may be considered correct in the absence of a parity-error indication. Therefore a carry from any position, say $k-1$, may be logically combined with previously checked functions to predict the presence or absence of a carry from the next higher-order bit position k. The logic used in this prediction represents a redundant serial-carry generation from each bit position to the next. By comparing predicted and actual carries for each bit position, the presence of any single error can be detected. Fig. 17 shows such an arrangement. Clearly, it is applicable only if the complete carry number is available. In the case of our particular design, the sum is generated in such a highly parallel way that a line representing the carry from each bit position does not appear within the logic. However, since

$$S_k = H_k \forall C_{k-1}$$

then

$$C_{k-1} = H_k \forall S_k$$
$$(H_k \forall S_k = H_k \forall H_k \forall C_{k-1} = C_{k-1})$$

and the presence of a carry from the $k-1$ position may therefore be determined from the sum and half-sum of the kth position. The checking logic which uses the carries thus deduced from the sum has the added advantage of checking the validity of the sum. Now, since

$$C_k = g_k + H_k C_{k-1}$$

substituting for C_{k-1} gives

$$C_k = g_k + H_k \cdot (H_k \forall S_k)$$

consequently

$$C_k = g_k + H_k \cdot \bar{S}_k$$

Thus a carry into the $k+1$ position may be predicted from the sum, half-sum, and carry-generate functions of the kth position. The logic for this arrangement is shown in Fig. 18. Fig. 19 shows the block diagram with carry checking added.

Actual Design

A number of control refinements were incorporated into our specific design. They serve to illustrate flexibility of the approach.

The low-order 6-bits of the adder consisted of two 3-bit groupings, and control lines were added in order that the 24-bit adder could also operate as independent 6-bit and 18-bit adders. In addition the low-order 6-bit portion was two emitter-follower stages shorter than the other 18 bits, so that the resultant sum became available to other parts of the computer at an earlier time within the add cycle. In the independent mode, either or both portions of the adder may receive a "plus one" on a control input for counting or for subtracting (in conjunction with complement input data). In the 24-bit mode, this same facility is of course available for the adder as a whole. A 6-bit overflow carry is developed for external use. This output is not affected by the split or combined mode of operation. Finally, a minus-one control line is provided for counting down. The operation is equivalent to subtraction by adding the complement of the binary value "1" to the "A" input along with a carry into the low-order position of the adder. This control line simulates the complement of "1" by conditioning the B data inputs for all except the low-order bit.

In general, these features were obtained by minor variations in the logic already described. Fig. 20 shows a block diagram of the design, including logical expressions for the added functions. The control lines for these added functions also appear on Fig. 19, which is the block diagram for the complete adder and checking system.

Reference

1. One Microsecond Adder Using One-Mc Circuitry, **A. Wienberger, J. L. Smith.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., June 1956, pp. 65–73.

Algorithms for Logical Design

ANN C. EWING
NONMEMBER AIEE

J. PAUL ROTH
NONMEMBER AIEE

ERIC G. WAGNER
NONMEMBER AIEE

A SERIES of "extraction algorithms" are presented that are applicable to the design of a wide class of combinational circuits. A special operation, the #-product, is introduced for carrying out these algorithms. Complete algorithms for the design of minimal and near-minimal 2-level AND-OR circuits are given, utilizing the extraction algorithms and the #-product. (These algorithms have been programmed for the International Business Machines Corporation (IBM) *704* and are available through the SHARE Library. Data are shown on the usage of these programs by IBM circuit designers.) Indication is also given as to how the extraction algorithms may be applied to a broader class of problems, namely to the design of disjunctions of nonsingular circuits.

The general extraction algorithm and the local extraction algorithm have been described in previous publications. The statement of these algorithms in terms of the #-product, and, especially, the data on the programs, however, are new with this paper.

Statement of the General Problem

Consider combinational circuits which have the property that they consist of a disjunction (OR-ing together) of some well-defined class of "primary circuits," as shown in Fig. 1. For example, in the case of 2-level AND-OR circuits the primary circuits are simply AND-circuits.

Let P be a set of primary circuits. Suppose that each member p of P has attached to it an integer called its cost $\mu(p)$. The cost of a disjunction M of a set of primary circuits is defined to be the sum $\mu(M)$ of the costs of the members of P in the disjunction. For example, for the normal form case it is possible to assign to each AND an integer equal to the number of its inputs; this gives minimal input-diode count as the criterion for circuit minimality.

Let a circuit have inputs $a_1, \ldots, a_n$ and (for now) but a single output. The desired operation of this circuit is described by a matrix (table) of 1's and 0's, to be called the "on-matrix" of the

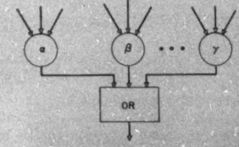


Fig. 1. A disjunction of primary circuits

Paper 59-1191, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9–14, 1960. Manuscript submitted June 15, 1959; made available for printing February 21, 1961.

Ann C. Ewing, J. Paul Roth, and Eric G. Wagner are all with the International Business Machines Corporation Research Center, Yorktown Heights, N. Y.

circuit. In particular, let the matrix be $m \times n$, each row corresponding to a combination of input signals for which it is desired that the circuit output be on (there being m such combinations in all), and each column corresponding to an input variable. For example with $n=3$, and inputs a_1, a_2, a_3, let the circuit be on for the set of four input combinations described by the rows of the following table:

INPUT		
a_1	a_2	a_3
on	off	off
off	on	off
on	on	on
off	off	off

then the matrix would be:

$$\begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

The order of the rows is immaterial. If, in addition, there are certain combinations of inputs for which it does not matter if the circuit is on or off, these input combinations in like manner form a matrix which is known as the "don't-care matrix" of the circuit. Let the matrix consisting of the rows of the on-matrix together with the rows of the don't-care matrix be called the "input matrix" of the circuit.

If a matrix of the type described contains two rows that are identical except in the ith co-ordinate, it is convenient to replace these rows by a single row with an x in the ith co-ordinate, standing for the fact that this co-ordinate can be either 1 or 0. A row with more than one x is then defined by induction. It is easily seen that a row $(xx\ldots x)$ of all x's corresponds to the set of all possible inputs. Note that there is now a one-to-one correspondence between such matrices, and the set of normal form Boolean expressions; e.g., the matrix

$$\begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{bmatrix} 1 & 0 & x \\ x & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

corresponds to the expression

$$b_1\bar{b}_2 \vee b_2\bar{b}_3 \vee \bar{b}_1\bar{b}_2\bar{b}_3$$

A matrix with no x's will be called a "canonical matrix" (corresponding to canonical normal form). Note that the on-, input-, and don't-care matrices are canonical. The rows of a canonical matrix will be referred to as its "vertices." It is clear that for a given matrix C^* of 1's, 0's, and x's there is a unique corresponding canonical matrix C. It is said then that C^* is a "perfect cover" of C. If C^{**} is a matrix whose corresponding canonical matrix contains all the rows of the canonical matrix C, then C^{**} is called a "cover" of C and also of C^* where C^* is any exact cover of C. The rows of a matrix shall sometimes be referred to as "cubes." The set of all cubes covered by a matrix K is called the (cubical) complex determined by K. (For the motivation behind this terminology see references 1–3.)

Let some set P of primary circuits be given with inputs from the set $a_1, \ldots, a_n$ of input variables. If p is an element of P, let $\Gamma(p)$ denote a cover of the on-matrix of p.

The general design problem solved by the extraction algorithms can then be stated as follows. Let covers be given of the input matrix K and the on-matrix L for a desired circuit, and also the set P of primary circuits with a cover $\Gamma(p)$ for every $p \epsilon P$. The circuit design problem is then to find a subset M of P such that (1) each vertex of L is covered by the on-matrix of at least one element of M, (2) K covers $\Gamma(M)$, and (3) $\mu(M)$ is a minimum. That is, the circuit consisting of the OR-ing together of the elements of M is such that (1) at least one element of M is on for each input in the on-matrix, (2) no element of M is on for an input that is not in the input matrix, and (3) the sum of the costs of the elements of M is a minimum. For the case of 2-level AND-OR circuits, P consists of all AND-circuits, and the cost is, say, the number of inputs. Then given input and on-matrices K and L, the problem in this case would be to find the disjunction of AND-circuits whose on-matrix contains L and is contained in K.

For purposes of ease of presentation a "covering-type" problem is defined. The general extraction algorithm is described for two variations of this problem.

An Abstract Covering-Type Problem

Rather than considering the problem of "extracting" M from P (which might well be infinite), consider the problem of extracting M from some suitable subset of $G(K)$, the set of all primary circuits whose on-matrices are covered by K. In the cases of interest, $G(K)$ is finite. An algorithm will be given in a subsequent section for generating a suitable subset of $G(K)$ for the case of 2-level AND-OR circuits. The problem of extracting M from $G(K)$ is approached in a very general way. To do this, the problem is restated in a general set-theoretic form (compare reference 3).

Let V and W be sets with V containing W. Let G be a finite set and Γ a map of G into the set of all subsets of V. If v belongs to the set $\Gamma(g)$ for g in G, then g is said to "cover v." A "cover" C of W is a set of members of G such that each member of W is covered by some member of C.

With each member g of G let there be associated a positive integer $\mu(g)$ called the "cost" of g. The cost of a cover C is the sum of the costs of its members.

The Problem

The problem is to find a cover M of W having a minimum cost. Letting V be the set of vertices of the input matrix K, W the set of vertices of the on-matrix L, and G the set of primary circuits such that $\Gamma(g)$ is covered by K for all g in G, it is seen that the abstract problem is the same as the circuit design problem described earlier. The advantage of the more general presentation of the problem is that such extraction problems show up in applications other than circuit design.

The General Extraction Algorithm

For this algorithm assume that the complete set of elements of $G(K)$ is given. The steps of the algorithm are as follows:

1. Let $W_1 = W$ and $G_1 = \{g \mid g \epsilon G$ and $\Gamma(g) \cap W \neq \phi\}$.

2. Let $g, g' \epsilon G_1$; it is said that "g' is less than g" (with respect to W_1) if $\Gamma(g') \cap W_1 \subseteq \Gamma(g) \cap W_1$ and $\mu(g') \geq \mu(g)$ and equality does not hold in both relations. If equality does hold in both relations, let one of g and g' be decreed by lexicographical order to be less than (with respect to W_1) the other. Let $g' \overset{<}{(1)} g$ denote that g' is less than g (with respect to W_1). The relationship $\overset{<}{(1)}$ is a partial order. If $g' \overset{<}{(1)} g$, then there exists a minimum cover not containing g', for $\Gamma(g)$ covers at least as much as $\Gamma(g')$ and g costs no more than g'. Members of G_1 which are maximal under $\overset{<}{(1)}$ are called "W_1-elementary." Let Z_1 denote the set of W_1-elementary members of G_1. The first step of the algorithm is to find Z_1.

3. A member e of Z_1 is a "W_1-extremal" if there is an element w of W_1 which is covered by e but by no other element of Z_1. That is, e is a W_1-extremal if:

$$[\Gamma(g) - (V - W_1) - \Gamma(Z_1 - \{e\})] \neq \phi$$

Let $E_1 = E(G_1, W_1)$ denote the set of W_1-extremals of Z_1. Clearly every minimal cover of W_1 contains E_1. Thus the second step of the algorithm is to find E_1. The pair (G_1, W_1) is said to be reducible if $E_1 \neq \phi$; otherwise irreducible. The reducible case will be treated first.

4. The case when (G_1, W_1) is reducible. Let $M_0=E_1$. Let $(G_1, W_1)\supset(G_2, W_2)$ mean that $G_1 \sqsupset G_2$ and $W_1 \sqsupset W_2$. Having formed pairs $(G_1, W_1)\sqsupset(G_2, W_2)\supset\ldots\supset(G_r, W_r)$ such that each pair is assumed to be reducible with $W_r\neq\phi$, it is possible to define for elements of G_r, the relation "less-than" (with respect to W_r) almost exactly as in step 2 with G_r and W_r replacing G_1 and W_1. The difference is that if equality holds in both relationships i.e., if $\Gamma(g')\cap W_r=\Gamma(g)\cap W_r$ and $\mu(g)=\mu(g')$ and if in the pair $(G_{r-1}, W_{r-1})g'$ was less than g (with respect to W_{r-1}), then let g' be less than g (with respect to W_r) also. If equality holds in both relationships, then let one be chosen by lexicographical order, to be less than the other (with respect to W_r). Let $g'\overset{<}{(r)}g$ denote that g' is less than g (with respect to W_r). Clearly $\overset{<}{(r)}$ is a partial order. Elements of G_r which are maximal under this partial order are said to be "W_r-elementary." Let

$$Z_r=Z(G_r, W_r)=\{g\in G_r | \text{ there does not exist } g'\in G_r, g\overset{<}{(r)}g'\}$$

denote the set of W_r-elementary members of G_r. A member e of Z_r is termed a "W_r-extremal" if there is an element w of W_r covered by e but by no other member of Z_r. Let $E_r=E(G_r, W_r)$ denote the set of W_r-extremals of G_r. By hypothesis $E_r\neq\phi$.

Let

$$M_r=M_{r-1}\cup E_r$$

$$W_{r+1}=W_r-\Gamma(E_r)$$

$$G_{r+1}=Z_r-\{g|g\in G_r \& \Gamma(g)\cap W_{r+1}=\phi\}$$

Proposition. Given the pair (G, W) and map $\Gamma:G\rightarrow W$ there exists a minimum cover containing M_r. For proof see reference 3.

Corollary. If $W_{r+1}=\phi$ then M_r is a minimum cover for the original problem.

Let (G, W) be a pair. Consider the sequence

$$(G, W)\supset(G_1, W_1)\supset\ldots\supset(G_r, W_r)\supset(\phi, \phi)$$

with

$$W_{i+1}=W_i-\Gamma(E_i)$$

$$G_{i+1}=Z(G_i, W_i)-\{g|\Gamma(g)\cap W_{i+1}=\phi\}$$

If for some i, $W_i\neq\phi$ but $W_{i+1}=\phi$, then (G, W) is said to have a terminating sequence and to be "completely reducible." The irreducible case will be considered next.

5. Case when (G, W) is not completely reducible. Suppose that the sequence associated with (G, W) does not terminate, that is, for some $r>0$, $W_r\neq\phi$ but $E_r=\phi$, so that $W_{r+1}=W_r-\Gamma(\phi)=W_r$.

The extraction algorithm treats this case by solving two problems, each simpler than the original. The procedure is termed "branching."

Branching. Let z_0 be an element of G_r. The first problem seeks a minimum cover which "contains" z_0; the second, one that does not. Whichever cover has lower cost is clearly a minimum for the original problem. The two problems are defined in parallel columns, the one on the left being the problem from which a minimum cover containing z_0 is sought. Let:

$$\begin{array}{l|l}
W^1_{r+1}=W_r-\Gamma(z_0) & W^0_{r+1}=W_r \\
G^1_{r+1}=Z(G_r, W_r)- & G^0_{r+1}=G_r-\{z_0\} \\
\qquad\{z\Gamma(z)\cap W_r=\phi\} & \\
M_r^1=M_{r-1}\cup\{z_0\} & M_r^0=M_{r-1}
\end{array}$$

The branching procedure changes the irreducible problem (G, W_r) into problems (G^0_{r+1}, W^0_{r+1}) and (G^1_{r+1}, W^1_{r+1}) where the cardinality of both G^0_{r+1} and G^1_{r+1} is less than that of G_r.

This completes the description of the general extraction algorithm. For a proof that the general extraction algorithm produces a minimum, see reference 3.

In the next section a variation on the extraction algorithm is introduced which requires considerably less than the generation of all the elementary members of G. It is in effect a kind of "localization" of the activity of the algorithm.

Some motivation concerning the efficiency of the extraction algorithm is now given. Perhaps this is most easily seen in the special case of minimizing over the class of 2-level AND-OR circuits. In this context the first-order extremals correspond to the set of AND-OR blocks which *must* belong to every minimum circuit. Now, after going through the first less-than operation, the second-order extremals are those blocks which *must* belong to every minimum circuit which does not contain the blocks eliminated through this less-than operation. A similar remark holds for all higher order extremals. It is the fact that the succession of choices of extremals of all orders are *forced* that gives the algorithm its power.

The Local Extraction Algorithm

Let N be a subset of V and let D be a subset of G for which $\Gamma(D)=N$. A "D-periphery" $P(g)$ of an element g of G is the set of all g' of G for which $[\Gamma(g')\cap\Gamma(g)]-\Gamma(D)$ is not empty. The significant case here is when $N=V-W$. In this case it can be noted that if g' is less than g, then of course g' must belong to the periphery of g. It is furthermore noteworthy that the determination of whether or not g is an extremal depends only on the elements in the periphery of g. It follows from these two observations that the extraction algorithm can be localized, in the sense that the computations can be conducted in terms of peripheries rather than in terms of G as a whole.

This possibility is especially advantageous in the case when G is very large but the given cover, or a cover generated therefrom, is small. This distinction is especially important when the algorithm is executed on a computer with relatively small memory capacity. In particular, when the algorithm is performed on a machine of relatively small memory capacity, this possibility could well determine whether or not a particular problem can be solved. A description of the algorithm will now be given. A member of definitions are first required. Let D be a "perfect" cover of $N=V-W$, that is, $\Gamma(D)=N$. The condition that g' be less than g (with respect to W) can be expressed by the conditions: $\Gamma(g')-\Gamma(g)-\Gamma(D)=\phi$, and $\mu(g)\leq\mu(g')$, with the same rules holding in the case of ties. Let $Z(g)$ denote the elements of the D-periphery $P(g)$ of g that are maximal with respect to $<$. It is very important to remember that the maximality may be conducted *exclusively* in the periphery.

A rephrasing of the definition of W-extremal, appropriate for the local algorithm, is: An element g of G is a W-extremal if $\Gamma(g)-\Gamma(Z(g)-\{g\})$ is not empty.

In terms of these definitions the local algorithm is given as follows: Let C be a cover of V, and D a cover of $N=V-W$. Replace a cover C by a cover C^* consisting of maximal elements: this can be accomplished by generating the periphery of each element of C and selecting therefrom a maximal element greater than the element of C. It is easy to see that all W-extreme elements must be contained in C^*. The inductive procedure necessary for the local algorithm is essentially the same as that used for the general algorithm. The branching aspect is similar, with one essential modification: When an element g^* of G is discarded as if it were less than another element of G, then g^* must be subtracted from every periphery. (There are actually several ways for accomplishing the equivalent of this.)

An Approximate-Minimum Algorithm

Originally designed as a method for providing an estimate of the running time for the extraction algorithm, the algorithm to be described has seen use as a method in itself for rapidly producing "near-minimal" solutions. The steps are as follows: Let C_0 and D_0 be covers of V and $N_1=V-W$.

1. Replace cover C_0 by a cover C_1 consisting of maximal elements and let $D_1=D_0$.

2. Find the set E_1 of 1-extremals. They are all in C_1 and may be located by the same test as in the local algorithm.

3. Let $C_2=C_2-E_1$; $D_2=D_1\cup E_1$; $A_2=E_1$.

4. Simple-order C_2 compatible with the partial order of increasing cost.

5. For each g in C_2 chosen according to the simple order, form $\Gamma(g)-\Gamma(D_2)$.

If this difference is empty for each g in C_2, then A_2 forms an approximately minimum cover (in this case A_2 is actually the unique minimum). Let g' be the first c in C_2 for which $\Gamma(g)-\Gamma(D_2)$ is not empty. Form

$$C_3=C_2-\{g'\},\ D_3=D_2\cup\{g'\},\ A_3=A_2\cup\{g'\}$$

and proceed with C_3, D_3, A_3 exactly as with C_2, D_2, A_2 above. If for each g in C_3 the $\Gamma(g)-\Gamma(D_3)$ is empty, then A_3 constitutes an approximate minimum, and if not, then one proceeds through another iteration, etc., and very quickly obtains an approximate minimum.

This completes the description of the approximate algorithm.

The #-Product

This section shows how these algorithms can be applied to the circuit design problems of the kind described in the introductory sections of this paper. In particular, methods are presented for carrying out the algorithms when V, W, and $\Gamma(g)$ for all $g \epsilon G$ are given by matrix covers, i.e., matrices of 1's, 0's, and x's. To do this an operation on pairs of matrices called the #-product (sharp product) is introduced.[1] It will be indicated how all the operations in the algorithms (outside of ordering and choice operations) can be performed by means of this #-product. In the following section it will be shown how this #-product can also be used to generate the set $G(K)$ and peripheries for the 2-level AND-OR circuit design problem.

The reader is cautioned that despite its universality the #-product is neither the only nor always the most efficient way to carry out some of these operations. References 1 and 2 give complete algorithms for the 2-level AND-OR circuit that do not include the #-product. Furthermore, the programs described in the final section of this paper use other special products as well as the sharp.

The #-product $c\#d$ of cubes c and d is a particular set of cubes forming a cover of the vertices of c which are not in d. This product is thus a species of differencing operation. Logically speaking, if c and d correspond to terms S and T respectively, then $c\#d$ is a way of forming $S\cdot\bar{T}$. It is defined explicitly by the following rules:

1. $c\#d=\phi$ if $c\subset d$.
2. $c\#d=c$ if $c\cap d=\phi$; i.e., $c_i\cap d_i=\phi$ for some i.
3. Otherwise, let $i(1), \ldots, i(r)$ be the set of co-ordinates $i(j)$ for which $c_{i(j)}=x$ and $d_{i(j)}=0$ or 1. Then $c\#d$ contains r cubes, the jth being $(c_1, \ldots, c_{i(j)-1}, \bar{d}_{i(j)}, c_{i(j)+1}, \ldots, c_n)$, which has all its co-ordinates equal to those of c, except for the $i(j)$th, which is 1 if $d_{i(j)}$ is 0 and 0 if $d_{i(j)}$ is 1.

Examples: $x110\#xx10=\phi$; $xx1x1\#x1x0x=x01x1+xx111$; $x11x0x\#1x100x=011x0x+1110x$; $xxx\#101=0xx+x1x+xx0$.

Machinewise, the first two operations are performed first and in parallel, that is, on all co-ordinates simultaneously.

Subsuming the Matrix

An operation on matrices, called subsuming, will now be given.[4] Let C be a matrix. Then subsuming C gives a new matrix consisting of those cubes c of C such that there is not another cube c' in C such that $c\#c'=\phi$, i.e., there is no other cube (row) in C that covers all the vertices of c. This trivial operation obviously results in a matrix that covers exactly what C covered and will do so without "simple" redundancies. Given a matrix C let $S(C)$ denote the result of the performance of the subsuming operation on C.

#-Product of Matrices

Next is defined the #-product of matrices A and B, each having n columns. Let $a\#B$ denote the iterated product of a, an element of A, with every element of B; by rule 1, this product is independent of the order selected for the product. Subsume the set of all such iterated products, $a\#B$. The result is the #-product $A\#B$ of matrices A and B.

Rendition of Set-Theoretic Operations

The set-theoretic operations $\supseteq$, $\cap$, employed in the general descriptions of the extraction algorithms can be restated in terms of #-products of matrices.

($-$). If A and B are matrices, then, as was shown before, $A\#B$ is a matrix covering those vertices in A that are not covered by B. But clearly then for g and g' elements of G, $\Gamma(g)\#\Gamma(g')$ corresponds to $\Gamma(g)-\Gamma(g')$.

($\supseteq$). If A and B are matrices and $A\#B$ is empty, then B covers every vertex of A. Thus, for g and g' elements of G, $\Gamma(g')\#\Gamma(g)=\phi$ corresponds to the relation $\Gamma(g)\supseteq\Gamma(g')$.

($\cap$). With matrices A and B, the product

$$(A\#((xx\ldots x)\#B))=(B\#((xx\ldots x)\#A))$$

is a matrix covering exactly those vertices that A and B have in common. In particular then for g and g' in G, the corresponding product $\Gamma(g)\#((xx\ldots x)\#1(g'))$ corresponds to the intersection $\Gamma(g)\cap\Gamma(g')$.

Rendition of General Extraction Algorithm

Let V, W, and $\Gamma(g)$ for all g in G be given by matrices. Then the basic operations in the general extraction algorithm can be rewritten as follows:

(1*). $W_1=W$, $G_1=\{g \epsilon G \mid (\Gamma(g)\#((xx\ldots x)\#W))\neq\phi\}$

(2*). $g' \overset{\leq}{(r)} g$ if and only if

$$[\Gamma(g')\#((xx\ldots x)\#W_r)]\#[\Gamma(g)\#((xx\ldots x)\#W_r)]=\phi$$

and $\mu(g')\geq\mu(g)$. Let Z_r denote the set of elements maximal under $\overset{\leq}{(r)}$.

(3*). An element of G is W_r-extreme if and only if

$$\Gamma(g)\#(V\#W)\#\Gamma(Z_r-\{e\})\neq\phi$$

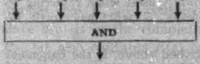
(4*). $M_r=M_{r-1}\cup E_r$ (with $M_0=\phi$)
$W_{r+1}=W_r\#E_r$
$G_{r+1}=Z_r-\{g \epsilon G_r \mid (\Gamma(g)\#((xx\ldots x)\#W_{r+1}))\neq\phi\}$

In like manner the basic definitions under "Local Extraction Algorithm" and "Approximate Minimum Algorithm" can also be rewritten in terms of the #-product.

Minimization Over the Class of 2-Level AND-OR Circuits

Correspondence Between 2-Level AND-OR Circuits and Matrices

First a correspondence will be established between 2-level AND-OR circuits, and matrices. Consider, for example, the AND-circuit:



This shall correspond to the cube $x110x01$. A set of such AND-circuits is represented by a matrix, each row of which is such a cube. In this context the set of AND-circuits considered will always be OR-ed together. To handle multiple-output 2-level AND-OR circuits, a device due to D. E Muller is used[6] (this is mentioned in a subsequent section).

In view of the correspondence between AND-OR circuits and matrices, it is possible to carry out the algorithm exclusively in terms of matrices and the #-operation. It may be mentioned that the cost assigned to a cube or to the AND-circuit which it represents need have only the single property that the cost *not* decrease as the number of inputs to the AND-block increases; for example, the cost might be 1 for each block.

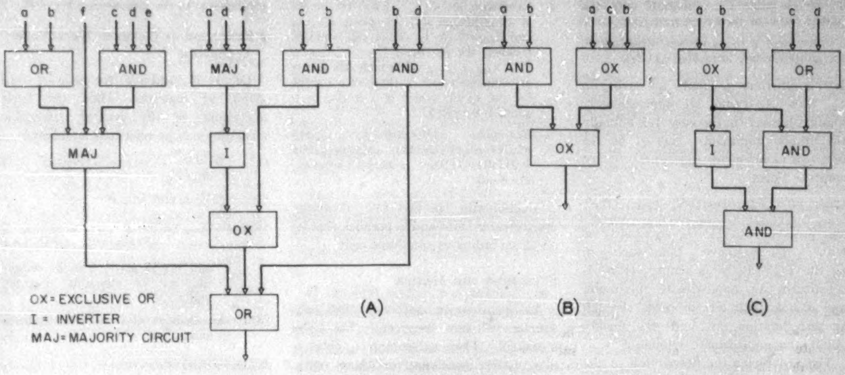


Fig. 2. Combination circuits

A—Disjunction of nonsingular circuits B—Not nonsingular: input "a" occurs twice C—Not nonsingular: the OX drives two circuits

THE MINIMIZATION PROBLEM

The specifications for the design of a 2-level AND-OR circuit are given by means of a so-called on-matrix W, which prescribes the set of input combinations for which the circuit must be on, and an input-matrix V which prescribes, in addition to W, the don't-care conditions, conditions for which it is immaterial whether the circuit is on or off.

To put the problem into the frame of the general covering problem:

1. G is the set of all cubes corresponding to AND-circuits that are on only for input combinations covered by the input-matrix V. In this context G will be termed a (cubical) complex. Consistent with the notation commonly used when referring to cubical complexes, G will henceforth be denoted by K.
2. V, as described above, is the canonical matrix determined by the input specifications; W is the subset of V consisting of input combinations for which the circuit must be on.
3. The mapping Γ is the map of a cube into its set of vertices, obtained by changing in all possible ways each x to a 1 or 0.

It must be emphasized that, with the $\#$-algorithm, it is not at all necessary to go to the canonical circuit specifications; everything may be done with covers.

COMPUTATION OF THE ELEMENTARY COCYCLES

An "elementary cocycle" is a cube belonging to K which is not covered by any other cube belonging to K. (An elementary cocycle corresponds to a prime implicant in the parlance of Quine.[4]) The restriction on the assignment of costs to AND-blocks, and hence to cubes, is such that if cube c' is covered by cube c, then the cost of c' is no more than that of c. Consequently, in order to find a minimum, it is possible to restrict attention to elementary cocycles, since clearly any minimum must be contained in this set. Let $Z(K)$ denote the set of elementary cocycles of K. The $\#$-method for computing the elementary cocycles of K will be given.

1. Let C be a cover of a complex K.
2. Form $(xx\ldots x)\#C=D$, subsuming at each stage after the first. D is the set of all elementary cocycles in $(xx\ldots x)\#C$.
3. Form $(xx\ldots x)\#D=F$, subsuming at each stage after the first. This product constitutes the set $Z(K)$ of all elementary cocycles of the complex K.

Lemma 1. If a, b are cubes, with n co-ordinates, the product $a\#b$ consists of the set of all elementary cocycles of the complex determined by the complement of b in a.

Lemma 2. If $a_1+\ldots+a_r$ forms the set of all elementary cocycles of a complex K and if b is a cube, then the product

$$(a_1+\ldots+a_r)\#b$$

contains the set of all elementary cocycles of the complex $K\#b$.

Proof of Lemma 2. Let $a_1+\ldots+a_r$ be the set of all elementary cocycles of K. Suppose that the product of this set of cubes with a cube b does not contain the cocycle z of $K\#b$, determined by the cubes of K that are not contained in b. Now since z is, by injection, also a cube of K, it must be covered by a cocycle of K, say by a_1. But z has empty intersection with b and thus of course lies in $a_1\#b$. By lemma 1, however, this product is the set of all cocycles of the complex

Table I. Problems Solved by MIN and MAP Programs

No. of Inputs	No. of Outputs	No. of Cubes in Input Matrix	No. of Diodes Saved	No. of Vertices	No. of Minutes of IBM 704 With Time MIN	With MAP
22	1	39	33	3,211,184	9	5
11	4	35	16	136	5	3
26	2	11	1	20,798,672	30	*
16	1	11	11	*	10	*
10	1	224	2,176	*	30	7
10	1	90	310	*	40	*

* Not available.

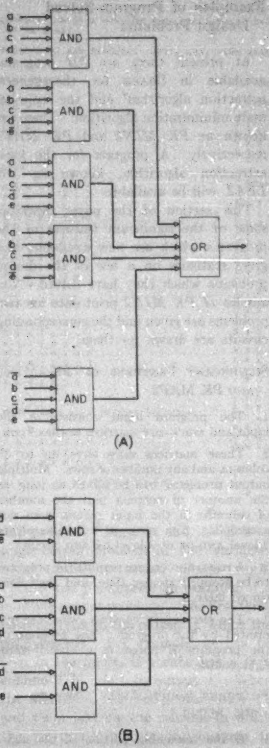


Fig. 3. Single-output-circuit problem

A—Before minimization
B—After minimization

determined by the cubes of a_1 not in b and hence one of these must contain z, which is a contradiction. (Q.E.D.)

Proof that #-Method Does Produce the Elementary Cocycles of K. It is only necessary to prove that $(xx\ldots x)\# C$ is the set of all cocycles of the complement of C. But this follows from lemma 2. (Q.E.D.)

With these tools the general extraction algorithm may easily be specialized to apply with effect to the 2-level AND-OR circuit problem. (An example is given in a later section.) In order to apply the abstract local extraction algorithm to this special problem, the only additional tool needed is to be able to generate peripheries efficiently. A method for doing this is described in the next section.

GENERATION OF PERIPHERIES FOR THE LOCAL EXTRACTION ALGORITHM

The procedure to be outlined requires at one point that a "cover" of the complement of a cover C of a complex K be computed. Now the #-product, as defined previously, is such that $xx\ldots x\# C$ produces the set of *all* cocycles of the complements. This, however, defeats the space-saving features that one desires from a local extraction algorithm. Thus a modified #-product, to be denoted #′, will be defined which produces merely a cover of the complement.

Procedure for Computing the Modified #-Product

1. Let C be a cover of K, $C=\{c^1, \ldots, c^p\}$ and let $D(0)=xx\ldots x$.

2. Form $D(0)\# c^1=D(1)$.

. . . .

$r+1$. The rth step of the procedure, $r<p$, yields a set $D(r)$ of cubes $D(r)=\{d_1, \ldots, d_{s(r)}\}$.
For each d_j of $D(r)$, form $d_j\# c_{r+1}=\{u_1, \ldots u_t\}=U(r+1, d_j)$.
From the set $U(r+1, d_j)$ a subset will be found which shall constitute an irredundant cover of $U(r+1, d_j)$.
Let

$D(r+1, 0)=D(r)$

Pick

$d_1 \epsilon D(r+1, 0)$

Form

$d_1\# c_{r+1}$

Let

$D(r+1,1)=(D(r+1,0)-d_1)\cup\{u \mid u \epsilon d_1\# c_{r+1}\}$

and

$u\#(D(r+1,0)-d_1)\neq\phi$

And, in general, for $j<s(r)$,

$$D(r+1,j)=(D(r+1,j-1)-d_j)\cup \\ \{u \mid u\epsilon d_j\# c_{r+1}\}$$

and

$u\#(D(r+1,j-1)-d_j)\neq\phi\}$

Finally

$D(r+1)=D(r+1, s(r))$

The final result is a set of cubes $D(p)$, where p is the number of cubes in C.

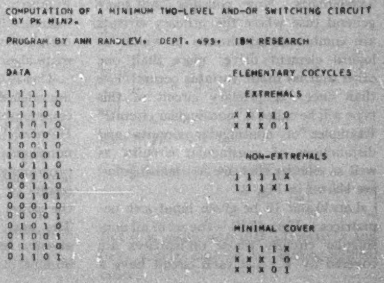
Lemma. $D(p)$ is a cocycle cover of $xx\ldots x\# K$, the complex in $xx\ldots x$ complementary to K.

Production of a Periphery.

Step 1. Let $D=\{d(1), \ldots, d(q)\}$ be a cover of the complement of K. Form $(xx\ldots x)\# d(1)$ and thence extract the subset with nonempty intersection with e; call this set $P_1(e)$. (These two operations can be conveniently combined into one.) This is the set of cocycles in the complex $(xx\ldots x)\# d(1)$ which intersect e.

Step $s+1$. Having formed the sth set $P_s(e)$, with s less than r, form the set $P_s(e)\# d(s+1)$ and then extract from this set the subset with nonempty intersection with e; call the resulting set $P_{s+1}(e)$.

The assertion is that the set $P_q(e)$, q being the number of elements in D, is precisely the periphery $P(e)$ of e. When D has exactly one element $d(1)$ then the complex K, the complement of the complex determined by D, has precisely the elements of the product $(xx\ldots x)\# d(1)$ as its set of elementary cocycles and the subset $P_1(e)$ of these which intersect with e is by definition the set of all cocycles contiguous to e. Suppose then that $P_{q-1}(e)$ is the set of all cocycles in the complex complementary to the set $d(1)\ldots d_{q-1}$ which are contiguous to e. This implies that $P_{q-1}(e)$ contains $P(e)$. By the fundamental property of the #-product,[1] the product $P_{q-1}(e)\# d(q)$ (here the #-product is performed on a subsuming-as-you-go basis) consists of the set of all cocycles of the complement of $d(q)$ in the complex determined by $P_{q-1}(e)$. Now the induction hypothesis implies that $P_{q-1}(e)$ contains $P(e)$. But since $d(q)$ is in the complement of $P(e)$, $P_{q-1}(e)$ also contains $P(e)$ and hence so does $P_q(e)$. Next one needs to show that $P_q(e)$ contains no more than $P(e)$. But this is not possible for the successive #-products with $d(1), \ldots, d(q)$ removed all vertices from the complement and the



```
COMPUTATION OF A MINIMUM TWO-LEVEL AND-OR SWITCHING CIRCUIT
 BY PK MIN2.

PRUGRAM BY ANN RANDLEV, DEPT. 493, IBM RESEARCH

DATA                                  ELEMENTARY COCYCLES

1 1 1 1 1                               EXTREMALS
1 1 1 1 0
1 1 1 0 1                             X X X 1 0
1 1 0 1 0                             X X X 0 1
1 1 0 0 1
1 0 0 1 0
1 0 0 0 1                               NON-EXTREMALS
1 0 1 1 0
1 0 1 0 1                             1 1 1 1 X
0 0 1 1 0                             1 1 1 X 1
0 0 1 0 1
0 0 0 1 0
0 0 0 0 1
0 1 0 1 0                             MINIMAL COVER
0 1 0 0 1
0 1 1 1 0                             1 1 1 1 X
0 1 1 0 1                             X X X 1 0
                                      X X X 0 1
```

Fig. 4. Computer print-out for single-output-circuit problem

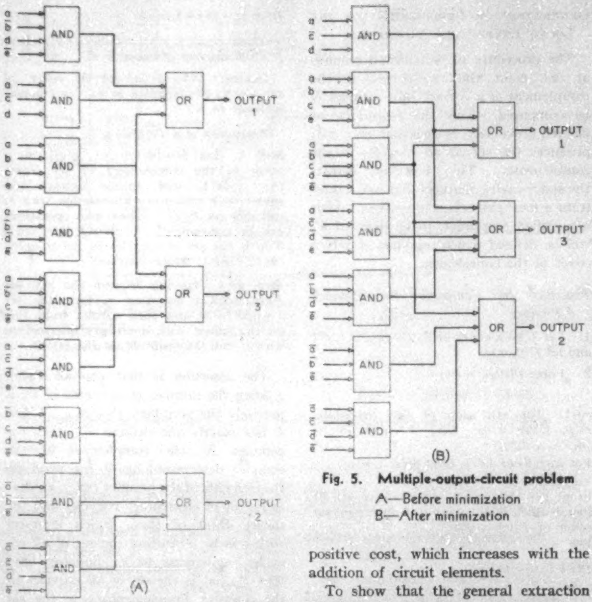


Fig. 5. Multiple-output-circuit problem

A—Before minimization
B—After minimization

only cubes of K that are retained are those which are contiguous. (Q.E.D.)

Thus a method for generating peripheries has been given and this is the key which leads to an efficient local extraction algorithm.

This completes the description of the algorithms for the case of 2-level AND-OR circuits. A description will now be given of a more general circuit design problem which can be solved by means of the extraction algorithms.

Nonsingular Circuits

In the case of 2-level AND-OR circuits the set of primary circuits contained only AND-circuits. Now consider the more general case where the primary circuits are combinational circuits in which no logical element drives more than one other and no input variable occurs more than once. A primary circuit of this type will be called a "nonsingular circuit." Examples of nonsingular circuits and disjunctions of nonsingular circuits, as well as circuits that are not nonsingular, are shown in Fig. 2.

Let V and W be given input-AND on-matrices. Here $G(V)$ is the set of all nonsingular circuits whose on-matrices are covered by V. Let each circuit have a positive cost, which increases with the addition of circuit elements.

To show that the general extraction algorithm applies, it is necessary to show that $G(V)$ can be reduced to a finite set.

First, any minimal circuit with n inputs contains only circuits in which there are no inverters in series. Second, the set T of all n-or-less input circuits without inverters-in-series is constructable by exhaustion. Given V with n columns, $G(V)$ can then be extracted from T by comparing the on-matrices of its elements with V, i.e.,

$$G(V)=\{t \mid t\epsilon T, \& \Gamma(t) \not\# V=\phi\}$$

Thus, since $G(V)$ can be generated, minimal disjunctions of nonsingular circuits can be designed.

Fortunately, far-better methods than the exhaustive one exist for generating $G(V)$; e.g., see references 3 and 6. Furthermore, the technique given in reference 3 can be employed to generate peripheries so that the local and approximate algorithms can also be employed.

This is all that can be said concerning nonsingular circuits without going into considerable detail as in reference 3. However, an experimental algorithm for nonsingular circuits is presently being programmed and when available the results of this program should indicate the power of the extraction algorithms. The next section reports on programs already completed for dealing with minimal 2-level AND-OR circuits.

Examples of Program-Solved Design Problems

At present there are *704* programs available in SHARE for the general extraction algorithm[7] and the approximate minimization algorithm. These are known as *PK MIN2* and *PK MAP2* respectively. A program for the local extraction algorithm, known as *PK LEA2*, will be available.

This section of the paper describes some of the significant features of the programs which are now available, and gives statistics on a few of the design problems which they have solved. Examples of *PK MIN2* print-outs for two problems are given and the corresponding circuits are drawn for them.

SIGNIFICANT FEATURES OF PK MIN2 AND PK MAP2

1. The program input consists of the input and don't-care matrices of the circuit.

2. These matrices may have up to 35 columns and any number of rows. Multiple output problems[8] can be solved as long as the number of outputs plus the number of columns in the input matrix does not exceed 35. The program will solve either 2-level AND-OR or 2-level OR-AND circuits.

3. A restart procedure permits the program to be stopped at any time, and then continued later.

4. The elementary cocycles computed are printed by the program. The progress of the program is noted in a special-sense light coding.

FEATURES WHICH APPLY ONLY TO PK MIN2

1. If the branching portion of the algorithm is employed, that part of the problem which requires its use is printed out by the program.

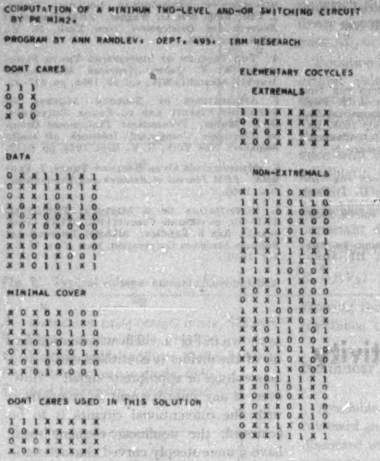
2. An approximate solution can be obtained at any time during the branching procedure.

STATISTICS ON PROBLEMS

Table I gives significant statistics on typical problems solved by *PK MIN* and *PK MAP* programs. These problems all arose in the design of IBM machines. The number of inputs is the same as the number of columns in the input matrix, and the number of vertices is the number of rows in the canonical form of the on-matrix. The algorithm, fortunately, does not use canonical form; the numbers show why!

EXAMPLES OF PROGRAM PRINT-OUTS

Examples of two small problems are given in this section. Diagrams of the circuit before and after minimization are given, as well as a program print-out of



```
COMPUTATION OF A MINIMUM TWO-LEVEL AND-OR SWITCHING CIRCUIT
 BY PK MIN2.

PROGRAM BY ANN RANDLEV.  DEPT. 493.  IBM RESEARCH

DONT CARES                                 ELEMENTARY COCYCLES

1 1 1                                         EXTREMALS
0 0 X
0 X 0                                      1 1 1 X X X X X
X 0 0                                      0 0 X X X X X X
                                           0 X 0 X X X X X
                                           X 0 0 X X X X X

DATA
                                              NON-EXTREMALS
0 X X 1 1 1 X 1
0 X X 1 X 0 1 X                            X 1 1 1 0 X 1 0
0 X X 1 0 X 1 0                            1 X 1 X 0 1 1 0
X 0 X X 0 1 1 0                            1 X 1 0 X 0 0 0
X 0 X 0 0 X X 0                            1 1 X 1 0 X 0 0
X 0 X 0 X 0 0 0                            1 1 X 1 0 1 X 0
X X 0 1 0 X 0 0                            1 1 X 1 X 0 0 1
X X 0 1 0 1 X 0                            X 1 X 1 1 1 X 1
X X 0 1 X 0 0 1                            X 1 1 1 1 X 1 1
X X 0 1 1 1 X 1                            1 1 X 1 0 0 0 X
                                           1 1 X 1 1 X 0 1
                                           X 0 X 0 X 0 0 0
MINIMAL COVER                              0 X X 1 1 X 1 1
                                           1 X 1 0 0 X X 0
X 0 X 0 X 0 0 0                            X 1 1 1 X 0 1 X
X 1 X 1 1 1 X 1                            X X 0 1 1 X 0 1
X X X 1 0 1 1 0                            X X 0 1 0 0 0 X
X X 0 1 0 X 0 0                            X X X 1 0 1 1 0
0 X X 1 X 0 1 X                            X X 0 1 0 X 0 0
X 0 X 0 0 X X 0                            X X 0 1 X 0 0 1
X X 0 1 X 0 0 1                            X X 0 1 1 1 X 1
                                           X X 0 1 0 1 X 0
DONT CARES USED IN THIS SOLUTION           X 0 X 0 0 X X 0
                                           X 0 X X 0 1 1 0
1 1 1 X X X X X                            0 X X 1 0 X 1 0
0 0 X X X X X X                            0 X X 1 X 0 1 X
0 X 0 X X X X X                            0 X X 1 1 1 X 1
X 0 0 X X X X X
```

Fig. 6. Computer print-out for multiple-output-circuit problem

the problem. Figs. 3 and 4 show a single-output problem in five variables which arose in actual design work. Figs. 5 and 6 comprise an example of a multiple-output problem. The first three columns refer to the output to which the cube belongs. For example, as can be seen in the figure, if a cube has a 0 in column 1 it refers to output 1. When the first three columns contain only 1's and x's or just x's, the columns in which the x's occur indicate the outputs to which the cube refers. Columns 4–8 are the inputs. All the output coding is added by the program as well as the don't-care terms which are used in the output system.

A New Minimization Program

PK MIN2 will be replaced soon by *PK MIN4* which is an improved version of the program. *PK MIN4* has all the features of *MIN2* plus several others. It uses a faster method in the extraction algorithm. Solutions are computed in *MIN4* by the following steps:

1. The elementary cocycles are computed by means of either the $*$- or $\#$-algorithms.
2. The elementary cocycles are sorted into groups of extremals or nonextremals by the $\#$-algorithm. (Let Z be the set of elementary cocycles and $z \epsilon Z$, then form the product $(z_i\#(Z-z_i))=R$. If $R=\phi$, then z_i is a nonextremal. Otherwise, z_i is an extremal.)
3. The nonextremals are then either eliminated or converted to higher order extremals. This section also uses the $\#$-algorithm in the following steps:

```
(a). Let $U$ equal the set of nonextremals and $u_i$ be an element of $U$, $i=1, \ldots, m$.
     Also let $E$ equal the set of extremals and $e^k$ be an element of $E$, $k=1, \ldots, n$.
     Let $i=k=0$.
(b). Add 1 to $k$ and choose $e_k$. If $k=n+1$, go to branching.
(c). Add 1 to $i$.
     If $i=m+1$, and $U=\{u_1, \ldots, u_m\}\neq\phi$, go to step (b).
     If $i\neq m+1$, go to step (d).
     If $U=\{u_1, \ldots, u_m\}=\phi$, the problem is solved.
(d). Form the product $u_i \cap e_k = t$.
     If $t=\phi$, go to step (c).
     If $t\neq\phi$, go to step (e).
(e). Form the product $u_i\#E=R$.
     If $R=\phi$, erase $u_i$ and go to step (c).
     If $R\neq\phi$, proceed to next step.
(f). Set $j=1$.
(g). If $j=i$, go to step (i).
     If $j\neq i$, proceed to next step.
(h). If $\mu u_i<\mu u_j$, go to step (j). Otherwise, go to step (i).
(i). Add 1 to $j$.
     If $j=m$, go to step (c).
     If $j\neq m$, go to step (g).
(j). Form the product $u_i\#(E\cup u_j)=S$.
     If $S=\phi$, erase $u_i$ and go to next.
     If $S\neq\phi$, go to step (i).
(k). Form the product $u_j\#(E\cup(U-u_j))=P$.
     If $P=\phi$, go to step (c).
     If $P\neq\phi$, add $u_j$ to the set $E$, add 1 to $n$, and remove $u_j$ from the set $U$.
     Then go to step (c).
```

4. If all nonextremals are not eliminated, then the branching routine is used. Essentially, branching uses steps 2 and 3 as subroutines.

The advantages of *MIN4* include:

1. A choice of algorithms for computing elementary cocycles.
2. The ability to have don't-cares associated with particular functions as well as with all functions in multiple-output problems.
3. The use of the $\#$-algorithm in the extraction algorithm.
4. The ability to minimize the complement of the original function.
5. Print-outs of the program's progress and statistics on the problem such as cost and number of branch levels.
6. Detailed print-outs of the program's progress at the request of the operator, by means of a sense switch.

Example

The following is an example of the minimization of a 2-level AND-OR circuit with an extraction algorithm.

Let

$$W=V=\begin{bmatrix}1&1&0&x\\x&1&1&x\\0&0&x&1\\0&x&1&0\\0&0&0&0\end{bmatrix}=W_1=G_1$$

then

$$(xxxx)\#W=\begin{bmatrix}1&0&x&x\\0&1&0&x\end{bmatrix}=W'$$

hence

$$Z_1=(xxxx)\#W'=\begin{bmatrix}1&1&x&x\\x&1&1&x\\0&x&1&x\\0&0&x&x\end{bmatrix}$$

is the matrix of cocycles; the first-order extremals are now found. It is seen that

$$1\ 1\ x\ x\#(Z_1-\{1\ 1\ x\ x\})=1\ 1\ 0\ x$$

$$x\ 1\ 1\ x\#(Z_1-\{x\ 1\ 1\ x\})=\phi$$

$$0\ x\ 1\ x\#(Z_1-\{0\ x\ 1\ x\})=\phi$$

$$0\ 0\ x\ x\#(Z_1-\{0\ 0\ x\ x\})=0\ 0\ 0\ x$$

thus

$$E_1=\begin{bmatrix}1&1&x&x\\0&0&x&x\end{bmatrix}$$

thus

$$M_1=\begin{bmatrix}1&1&x&x\\0&0&x&x\end{bmatrix}$$

and

$$W_2=W^1\#E_1=0\ 1\ 1\ x$$

$$G_2=Z_1-\{z\epsilon Z_1 \mid z\#((xxxx)\#W_2)\neq\phi\}=\begin{bmatrix}x&1&1&x\\0&x&1&x\end{bmatrix}$$

But it is easily seen that $x11x<0x1x$ and $0x1x<x11x$ and so by lexicographical order,

$Z_3 = x\ 1\ 1\ x$

Hence

$E_3 = x\ 1\ 1\ x$

so

$$M_2 = \begin{bmatrix} 1 & 1 & x & x \\ 0 & 0 & x & x \\ x & 1 & 1 & x \end{bmatrix}$$

and

$W_3 = \phi$

so M_2 is a solution. Using the number of inputs to the AND-circuits as the cost criteria, it is seen that the initial circuit had cost 15 and the final result cost 6.

References

1. Algebraic Topological Methods for the Synthesis of Switching Systems—I, J. Paul Roth. *Transactions*, American Mathematical Society, Providence, R. I., vol. 88, July 1958, pp. 301–26; also *ECP 56-02*, Institute for Advanced Study, Princeton University, Princeton, N. J., Apr. 1956.

2. Algebraic Topological Methods for the Synthesis of Switching Systems—II, J. Paul Roth. *Annals of the Computation Laboratory*, Harvard University, Cambridge, Mass., vol. 29, 1959, pp. 57–73.

3. Algebraic Topological Methods in the Synthesis of Switching Systems; III—Minimization of Nonsingular Boolean Trees, J. Paul Roth, Eric G. Wagner. *IBM Journal of Research and Development*, New York, N. Y., vol. 4, Oct. 1959.

4. The Problem of Simplifying Truth Functions, W. V. Quine. *American Mathematical Monthly*, Menasha, Wis., vol. 62, 1955, pp. 627–31.

5. Applications of Boolean Algebra to Switching Circuit and to Error Detection, D. E. Muller. *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, New York, N. Y., Sept. 1954, pp. 6–12.

6. Minimization Over Boolean Trees, J. Paul Roth. *IBM Journal of Research and Development*, Dec. 1960.

7. Computation of a Minimum Two-Level AND-OR Switching Circuit [PK MIN2 (Program)], Ann E. Randlev. SHARE, International Business Machines Corporation, New York, N.Y., Apr. 1959.

Design of Diode-Switch High-Sensitivity Trigger Circuits

MIRKO M. VOJINOVIC
NONMEMBER AIEE

SOLUTION OF MANY switching problems is facilitated by trigger circuits with lower triggering threshold energy requirements than are usual in conventional monostable multivibrator circuits. Important applications are to be found, for instance, in pulse-height analyzers and similar circuits for measurements in nuclear research, time modulation, and binary memory systems.

Even special components and careful design have not secured appreciably improved circuit performance by overcoming natural stability limitations. Increased switching efficiency and development of some new circuits, however, have emerged with use of transistors and other recently devised components. Major improvements have resulted from employing switching diodes in vacuum-tube multivibrator circuits; these have been described by Kandiah and Vojinovic in earlier works.[1,2] This paper is concerned with a more general treatment of the switching-diode technique in high-stability trigger circuits.

Diode-Switch Modification in Multivibrator Circuits

The driven transition of the conventional multivibrator is initiated when the loop gain value is equal to, or higher than, unity. For this condition, the so-called "nonconducting" tube in the stable, or quiescent, state should be brought to the region in which its transconductance is high. Stability requirement for the conventional circuit, however, demands that the quiescent state shall be remote from the unity-gain condition. These somewhat conflicting precepts nevertheless clarify the point that curvature of the tube characteristic will limit the sensitivity of the conventional circuit.

To prevent oscillations while maintaining normal tube currents, the loop gain may be reduced below unity by insertion of a suitable switching element at some point in the feedback-loop path. The properties of this element should be such that the loop gain of the circuit can be triggered to a value greater than unity by an appropriate signal.

A suitable switching network can be realized in a voltage divider, consisting of an ordinary resistance and a voltage-dependent, resistive, nonlinear element, as shown in Fig. 1. The attenuation factor of the divider is controlled easily by a bias voltage or appropriate signal. However, if any improvement in comparison with the conventional circuits is to be achieved, the nonlinear element must have a more steeply curved characteristic than a vacuum tube. The broken-line characteristic in Fig. 2(E) would offer the best theoretical possibilities for attaining good stability control and very high triggering sensitivity. In practice, the thing sought is a nonlinear element whose characteristic comes closest to the ideal one and, at the same time, is most suitable for use in a multivibrator circuit.

As an example, consider the conventional plate-to-grid coupled monostable multivibrator in Fig. 3(A), designed to be triggered by a positive input pulse. The triggering sensitivity could be increased by letting both valves *V1* and *V2* conduct normal currents and by inserting the previously described voltage divider (Fig. 1) in the feedback-loop path. There are two possibilities for inserting the voltage divider, as shown in Figs. 3(B) and 3(C). The properties of both circuits normally depend on the tubes and components, but most important are the characteristics of the nonlinear element itself.

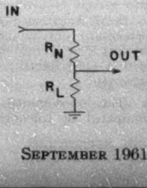
The stability criterion is given by the loop gain

$$G_L = GF \tag{1}$$

in which G = the product of the gains of the tubes *V1* and *V2*, and F = the attenuation factor in the feedback-loop path.

The circuit will remain in the stable state if

Fig. 1. Voltage divider, consisting of ordinary linear resistance R_L and voltage-dependent resistive nonlinear element R_N



Paper 61-99, recommended by the AIEE Electronics Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted August 18, 1958; made available for printing December 29, 1960.

Mirko M. Vojinovic is with the Institute of Nuclear Sciences "Boris Kidrich," Belgrade, Yugoslavia.

The author wishes to thank Dr. E. M. Williams and Dr. A. G. Milnes of the Carnegie Institute of Technology, Pittsburgh, Pa., for discussions and help in carrying out this work.

The paper was prepared by Mr. Vojinovic while he was serving as visiting Research Engineer in Electrical Engineering at Carnegie Institute of Technology, on leave from the Institute of Nuclear Sciences "Boris Kidrich," which organization he has now rejoined.

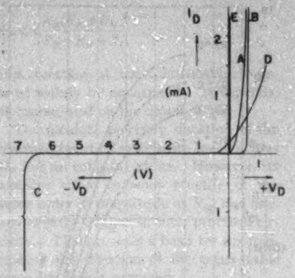


Fig. 2. Typical voltage–current characteristics for

A—Germanium point-contact diode, 1N34A
B—Silicon junction diode, 1N485A
C—Voltage-regulator silicon diode, SV-806
D—Thermionic vacuum diode, 6H6
E—Ideal broken-line characteristic

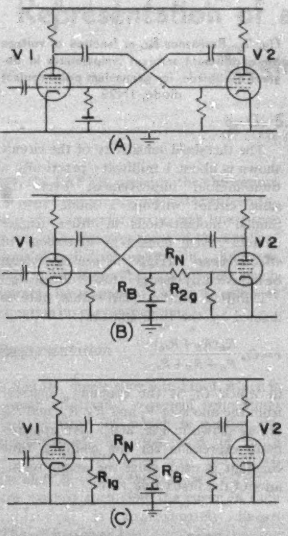


Fig. 3. Conventional plate-to-grid coupled monostable multivibrator designed for positive input pulse (A), and its sensitive versions, (B) and (C)

$$F<\frac{1}{G} \tag{2}$$

Suppose that the attenuation is obtained from a divider of the type described, having an attenuation factor

$$F=\frac{R_L}{R_N+R_L} \tag{3}$$

in which R_N = the resistance of the nonlinear element, and R_L = the fixed resistance [R_{1g} or R_{2g} in Fig. 3(B) or 3(C)].

Equation 3 makes clear that R_N must exhibit values high enough to satisfy equation 2. Moreover, to achieve a high stability and sensitivity at the same time, the rate of change of the nonlinear resistance with voltage should be as high as possible in the range to be used in the circuit. Thus, the applicability of a particular nonlinear element as well as its important design data should be apparent from a curve of

$$R_N=f(V_N) \tag{4}$$

showing the dependence of its nonlinear resistance on voltage.

Nonlinear Elements

Available nonlinear elements will be reviewed and their appropriate properties discussed with regard to their suitability for sensitive trigger circuits.

In principle, a vacuum diode might serve as a nonlinear element in the feedback loop. However, the vacuum diode has many disadvantages for such use. The sharpness of the characteristic is far from ideal, as shown in Fig. 2(D), stray capacitances are high, and heater voltage and considerable space for the tube itself must be supplied. Most of all, drift occurs as a consequence of heater-voltage fluctuations and other less controllable conditions.

Clearly, semiconductor diodes surpass vacuum diodes. However, it is necessary to differentiate among the wide variety of available semiconductor devices to determine their practicability in the feedback loop of sensitive trigger circuits.

Fig. 2 shows three different typical semiconductor-diode characteristics. Curve *A* represents a voltage–current characteristic of a germanium point-contact diode. Curve *B* describes an ordinary silicon junction diode, and curve *C* shows the same relation for a silicon voltage-regulator diode. All three resemble much more closely the ideal broken-line characteristic than do curves of the vacuum diode. In the case of the third diode, it seems that as an alternative to the forward region, namely, the avalanche region, could also be successfully used, especially because of improved temperature stability and better transient attributes in the reverse direction.

Since it is advisable to use the diode in the region where high rate of change of the resistance exists, the curves in Fig. 4 for the diodes in Fig. 2, defining the resistance–voltage relationship, are of special interest. The silicon junction diode *B* exhibits a high rate of change of resistance versus bias voltage and, at the same time, much larger resistance range than the germanium point contact diode *A*. However, for use in a sensitive multivibrator, the range from a few thousand ohms to a few megohms is sufficient and, in this region, the advantage of a silicon junction diode is not so apparent. The avalanche characteristic *C* in the useful resistance range has a lower rate of change than diodes *A* and *B*, and resistances smaller than 1,000 ohms are too small to

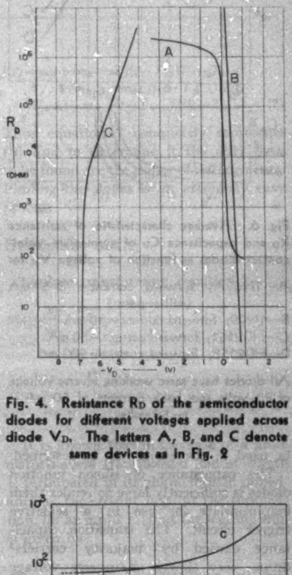


Fig. 4. Resistance R_D of the semiconductor diodes for different voltages applied across diode V_D. The letters A, B, and C denote same devices as in Fig. 2

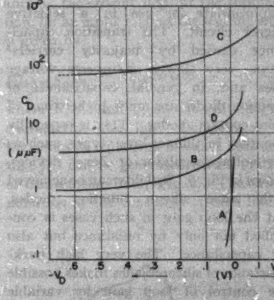


Fig. 5. Capacitance of diode C_D as a function of applied voltage V_D for

A—Germanium point-contact diode, 1N34A
B—Silicon junction diode, 1N485A
C—Voltage-regulator diode, SV-806
D—Silicon junction diode, 629

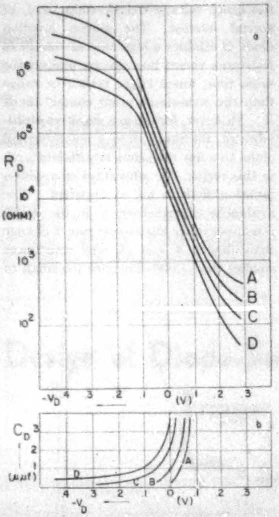


Fig. 6. Average characteristic of resistance R_D and capacitance C_D of germanium point-contact diodes as function of voltage V_D for

A—1N67A; forward current = 3–5 mA (milliamperes)
B—1N99; forward current = 10 mA
C—HD2151; forward current = 50 mA
D—HD2158; forward current = 200 mA

All diodes have same working inverse voltage of 80 volts and back currents of 50 mA at −50 volts

be used in the circuit to be described.

The capacitance of silicon junction diodes is sufficiently large to render them inappropriate for use in a sensitive trigger circuit. The transition capacitance caused by majority carriers[3] drops very slowly as inverse voltage rises and, in general, capacitances of junction diodes are much higher than of point-contact diodes. This is especially apparent in thinner-layer types, such as the voltage-regulator or Zener devices, shown in Fig. 5. Stability can be achieved with some silicon junction diodes, but the loop gain in such cases is controlled not only by resistance but also by capacitance. The very high back-resistance of silicon diodes makes possible the control of loop gain by variable capacitance alone, but the rate of change of this capacitance with voltage is not high enough for ordinary requirements and the circuits do not show particularly desirable properties.

Hence, the most convenient diodes for series nonlinear elements in sensitive multivibrators are the point-contact types. Further sifting of facts will reveal the most suitable point-contact diode for this purpose. Starting with the preference for diodes exhibiting higher back-resistances, consideration is given to four point-contact germanium crystal diodes, all having the same construction; all designed for the same inverse voltage; all with similar back-resistance at the specified voltage; differing only in the amount of forward current.

The attributes of these diodes, as shown in Fig. 6(A), show a similar slope for the straight-line portion, corresponding to a high rate of change of the resistance with voltage. Moreover, all straight-line portions cover, roughly, resistances of three orders of magnitude. The range of high rate of change does not cover the same values for different diodes, and the diodes designed for smaller currents are more suitable. There is an additional point in their favor, also; Fig. 6(B) shows they will control the circuit with the smallest possible interference from the capacitance of the diode.

Although semiconductor diodes have an advantage over thermionic diodes in that no variations of characteristics caused by heater supply voltage fluctuation can occur, semiconductor diode properties depend somewhat critically on ambient temperature, as shown in Fig. 7. For maximum sensitivity and reliability, therefore, temperature regulation is necessary in applications having varying ambient temperature, or compensation must be made for such changes.

Experimental Results

Although, clearly, the diode-switch modification may be applied to a wide variety of multivibrator circuits, this discussion will be limited to a single example, for sake of brevity.

Consider the plate-to-grid coupled monostable multivibrator of Fig. 3(B), with the nonlinear element between the anode of the first valve *V1* and the control grid of the second valve *V2*. Suitable component values for this circuit are shown in Fig. 8. The total gain is approximately 30. The voltage divider, consisting of R_{2g} and R_D, must make the loop gain <1 for stability. It can be shown from equations 2 and 3 that $R_N >$ 145,000 ohms. From the typical diode characteristics, the resistances are found to be about 145,000 ohms in the range where the rate of change of resistance R_D is fairly high and the diodes are, accordingly, suitable.

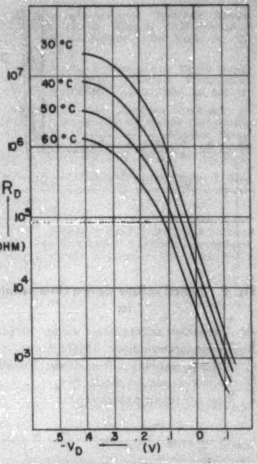


Fig. 7. Resistance R_D as function of voltage V_D at different ambient temperatures in degrees centigrade, for germanium point-contact diode, 1N99

The threshold sensitivity of the circuit shown is about 1 millivolt—practically a thousandfold improvement over the same circuit without a diode switch. Similar modifications of other trigger circuits reveal consistent extensions of effectiveness. Triggering sensitivity can be adjusted by changing bias voltage V_B.

Duration of the quasi-stable state is defined by the time constant

$$\tau = C_{1c}\frac{R_B(R_N+R_{2g})}{R_B+R_N+R_{2g}} \tag{5}$$

in which C_{1c} is the coupling capacitor from the anode of V_1 and R_B, R_N, and R_{2g} are as shown in Fig. 3(B). An approximate expression for the coupling time constant is used, since the anode resistance of the first valve *V1* is

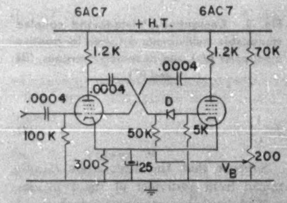


Fig. 8. Sensitive plate-to-grid coupled monostable multivibrator designed for positive input pulses

$$R_{1g} << \frac{R_B(R_N+R_{2g})}{R_B+R_N+R_{2g}}$$

and duration of the quasi-stable is defined mainly by equation 5. It depends, of course, also on the cutoff of *V*2.

To calculate precisely duration of the quasi-stable state is difficult, inasmuch as R_N has no constant value. However, its average value becomes roughly of the same order of magnitude as R_{2g}, and usually slightly smaller, in most practical circuits. This provides a basis for approximating the duration of the quasi-stable state for design purposes.

Conclusions

A thermionic-tube multivibrator circuit with a diode switch has been described. Modification of a conventional multivibrator by insertion of diodes results in improving triggering sensitivity by approximately a thousandfold. Essentially, the switching network consists of a voltage divider, in which a diode forms one element, and the dividing ratio is controlled by an appropriate signal. The diode should be of low capacitance and high back-resistance; in these respects, point-contact diodes have been found most suitable. The described diode-switch modification for improvement of triggering sensitivity may be applied to any conventional trigger circuit.

References

1. A Sensitive Pulse Trigger Circuit with a Stable Threshold, K. Kandiah. *Proceedings*, Institution of Electrical Engineers, London, England, vol. 101, pt. II, 1954, p. 239.

2. Increased Sensitivity of Trigger Circuits, M. Vojinovic. *Bulletin*, Department of Electronics, Institute of Nuclear Sciences "Boris Kidrich," Belgrade, Yugoslavia, vol. 8.

3. Electronic Semiconductors (book), E. Spenke. McGraw-Hill Book Company, Inc., New York, N. Y., 1958, p. 107.

A Note on the Approximate Representation of a Curve with Linear Segments

S. G. S. SHIVA
ASSOCIATE MEMBER AIEE

THIS PAPER presents the theoretical aspects of the approximate representation of an irregular curve with linear segments and discusses some of the applications of such a representation. An irregular curve, as used here, means one which cannot be expressed with a compact equation on easy inspection.

Representation

Let the curve shown in Fig. 1, which is chosen arbitrarily, represent some function of time $f(t)$ which is continuous, single-valued, bounded, and of finite duration T_n, $f(0)$ being zero and $f(t)=0$ for all $t<0$. Then the timings $T_1, T_2, \dots$ are marked such that the curve in each subinterval is approximately linear. Then $f(t)$ is given by:

$$\begin{aligned}f(t)=m_1t-m_1tU(t-T_1)+{}&(m_2t+K_2)U(t-T_1)-\\&(m_2t+K_2)U(t-T_2)+\dots+\\&(m_nt+K_n)U(t-T_{n-1})-\\&(m_nt+K_n)U(t-T_n)\end{aligned}\qquad(1)$$

where m_r is the slope of the segment between T_{r-1} and T_r, and K_r is the intercept this segment makes, when extended left, on the $f(t)$ axis. Rearranging the terms in equation 1,

$$\begin{aligned}f(t)=m_1t+{}&(m_2t+K_2-m_1t)U(t-T_1)+\\&(m_3t+K_3-m_2t-K_2)U(t-T_2)+\\&\dots+(m_nt+K_n-m_{n-1}t-K_{n-1})\\&U(t-T_{n-1})-(m_nt+K_n)U(t-T_n)\end{aligned}\qquad(2)$$

or

$$\begin{aligned}f(t)=m_1t+{}&[(m_2-m_1)(t-T_1)+K_2+\\&(m_2-m_1)T_1]U(t-T_1)+\\&[(m_3-m_2)(t-T_2)+K_3-K_2+\\&(m_3-m_2)T_2]U(t-T_2)+\dots+\\&[(m_n-m_{n-1})(t-T_{n-1})+\\&K_n-K_{n-1}+(m_n-m_{n-1})T_{n-1}]\\&U(t-T_{n-1})-(m_nt+K_n)U(t-T_n)\end{aligned}\qquad(3)$$

Reference to Fig. 1 shows that

$$K_r+m_rT_{r-1}=f(T_{r-1})\qquad(4)$$

and

$$K_{r-1}+m_{r-1}T_{r-1}=f(T_{r-1})\qquad(5)$$

Using equations 4 and 5 in 3:

$$\begin{aligned}f(t)=m_1t+{}&(m_2-m_1)(t-T_1)U(t-T_1)+\\&(m_3-m_2)(t-T_2)U(t-T_2)+\dots+\\&(m_n-m_{n-1})(t-T_{n-1})U(t-T_{n-1})-\\&(m_nT+K_n)U(t-T_n)\end{aligned}\qquad(6)$$

If the last segment, the one between T_{n-1} and T_n, is assumed to be almost coincident with the t axis so that $m_n\approx 0$ and $K_n\approx 0$, then equation 6 reduces to:

$$\begin{aligned}f(t)=m_1t+{}&(m_2-m_1)(t-T_1)U(t-T_1)+\\&\dots+(m_{n-1}-m_{n-2})(t-T_{n-2})U(t-\\&T_{n-2})-m_{n-1}(t-T_{n-1})U(t-T_{n-1})\end{aligned}\qquad(7)$$

Thus equation 7 completely represents $f(t)$ and to determine it the only data to be found are the slopes of the segments. Finding the slopes is an extremely easy process.

Generation of f(t)

Laplace transforming equation 7:

$$\begin{aligned}F(s)=\frac{m_1}{s^2}+\frac{m_2-m_1}{s^2}e^{-sT_1}+\dots+{}&\\\frac{m_{n-1}-m_{n-2}}{s^2}e^{-sT_{n-2}}-{}&\\\frac{m_{n-1}}{s^2}e^{-sT_{n-1}}&\end{aligned}\qquad(8)$$

It is known that in the case of a finite distortionless transmission line of length 1, terminated in its characteristic impedance z_0, the output $F_0(s)$ at the termination, for an input $F_i(s)$ at the sending end, is given by:

$$F_0(s)=F_i(s)e^{-1R\sqrt{\frac{L}{C}}}e^{-1s\sqrt{LC}}\qquad(9)$$

where R, L, and C, all per-unit length, characterize the line. Comparison of the terms of equations 8 with 9 indicates that $f(t)$ can be generated with a system of transmission lines; Fig. 2. The determination of the constants of each line is

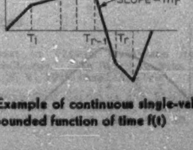


Fig. 1. Example of continuous single-valued bounded function of time f(t)

Paper 61-804, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted January 12, 1961; made available for printing April 18, 1961.

S. G. S. Shiva lives in Bangalore, India.

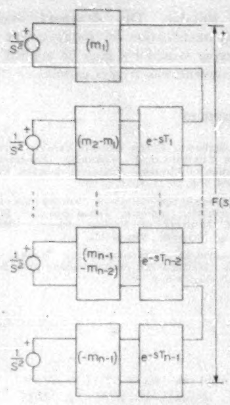


Fig. 2. System of transmission lines

straightforward. Taking some terms in equation 8 at random:

$$|m_k-m_{k-1}|=e^{-1_kR_k\sqrt{C_k/L_k}}$$

$$\text{for } |m_k-m_{k-1}|<1 \qquad (10)$$

and

$$1_k\sqrt{L_kC_k}=T_k \qquad (11)$$

In equations 10 and 11 (m_k-m_{k-1}) and T_k are known and 1_k, R_k, L_k, and C_k are to be determined. Introducing equation 11 in equation 10,

$$e^{-(R_k/L_k)T_k}=|(m_k-m_{k-1})| \qquad (12)$$

From equation 12 the ratio (R_k/L_k) is fixed. Clearly there are innumerable combinations of R_k, C_k, L_k, and 1_k which will satisfy equations 10, 11, and 12. The one consistent with practical convenience is to be chosen. It should be pointed out that generation of $f(t)$ by the method indicated here is possible only if the duration of $f(t)$ is of such a magnitude that the corresponding time delay can be achieved practically. For the case $|m_k-m_{k-1}|>1$ some sort of amplifier and lossless transmission line can be used.

Conversion of One Function Into Another

Suppose some function $g(t)$ of duration T_g is to be converted into another function $h(t)$ of duration T_h, both the functions satisfying all the restrictions mentioned in connection with $f(t)$. If u's and v's are the slopes of the segments of $g(t)$ and $h(t)$ respectively, then, on the basis of equation 8:

$$\frac{H(s)}{G(s)}=\frac{v_1+(v_2-v_1)e^{-ST}1+\ldots+(v_{h-1}-v_{h-2})e^{-ST}h-2-v_{h-1}e^{-ST}h-1}{u_1+(u_2-u_1)e^{-ST}1+\ldots+(u_{g-1}-u_{g-2})e^{-ST}g-2-u_{g-1}e^{-ST}g-1} \qquad (13)$$

in which, T_1, T_2, … are assumed to be the same for both $g(t)$ and $h(t)$. If, in general:

$$T_k=kT_1 \qquad (14)$$

$$T_1=T \qquad (15)$$

and

$$e^{-ST}k=\alpha^k \qquad (16)$$

then equation 13 can be rewritten as

$$\frac{H(s)}{G(s)}=\frac{v_1+(v_2-v_1)\alpha+\ldots+v(v_{h-1}-v_{h-2})\alpha^{h-2}-v_{h-1}\alpha^{h-1}}{u_1+(u_2-u_1)\alpha+\ldots+(u_{g-1}-u_{g-2})\alpha^{g-2}-u_{g-1}\alpha^{g-1}} \qquad (17)$$

There are two cases to be considered.

Case 1

Suppose $T_h>T_g$, or $h>g$. Under this condition equation 17 can be written:

$$\frac{H(s)}{G(s)}=a(\alpha)+\frac{b(\alpha)}{c(\alpha)} \qquad (18)$$

where $a(\alpha)$, $b(\alpha)$, and $c(\alpha)$ are polynomials of $(h-g)$, $(g-1)$ and g degrees respectively. Let

$$a(\alpha)=a_0+a_1\alpha+\ldots+a_{h-g}\alpha^{h-g} \qquad (19)$$

$$b(\alpha)=b_0+b_1\alpha+\ldots+b_{g-1}\alpha^{g-1} \qquad (20)$$

$$c(\alpha)=c_0+c_1\alpha+\ldots+c_g\alpha^g \qquad (21)$$

Suppose $c(\alpha)$ is factored so that

$$\begin{aligned}c(\alpha)&=C(\alpha_1-\alpha)(\alpha_2-\alpha)\ldots(\alpha_g-\alpha)\\&=C\alpha_1\alpha_2\ldots g(1-\alpha/\alpha_1)(1-\alpha/\alpha_2)\\&\qquad\ldots(1-\alpha/\alpha_g)\end{aligned} \qquad (22)$$

Using equations 19–22 in equation 18:

$$\begin{aligned}H(s)=G(s)[&a_0+a_1\alpha+\ldots+a_{h-g}\alpha^{h-g}+\\&(b_0+b_1\alpha+\ldots+b_{g-1}\alpha^{g-1})C^{-1}\alpha_1^{-1}\\&\ldots\alpha_g^{-1}(1-\alpha/\alpha_1)^{-1}\ldots\\&(1-\alpha/\alpha_g)^{-1}]\end{aligned} \qquad (23)$$

In equation 23, the expression

$$(b_0+b_1\alpha+\ldots+b_{g-1}\alpha^{g-1})C^{-1}\alpha_1^{-1}\ldots\alpha_g^{-1}(1-\alpha/\alpha_1)^{-1}\ldots(1-\alpha/\alpha_g)^{-1}$$

can ultimately be reduced to the form

$$d_0+d_1\alpha+d_2\alpha^2+\ldots$$

Therefore equation 23 can be, finally, in the form:

$$\begin{aligned}H(s)&=G(s)[a_0+a_1\alpha+\ldots+a_{h-g}\alpha^{h-g}+d_0+d_1\alpha+d_2\alpha^2+\ldots]\\&=G(s)[(a_0+d_0)+(a_1+d_1)\alpha+\ldots+(a_{h-g}+d_{h-g})\alpha^{h-g}+\\&\qquad d_{h-g+1}\alpha^{h-g+1}+\ldots]\end{aligned} \qquad (24)$$

Using equation 16 in equation 24,

$$\begin{aligned}H(s)=G(s)[&(a_0+d_0)+(a_1+d_1)e^{-ST}+\\&\ldots+(a_{h-g}+d_{h-g})e^{-ST(h-g)}+\\&d_{h-g1+}e^{-ST(h-g+1)}+\ldots]\end{aligned} \qquad (25)$$

Thus $h(t)$ can be realized with a system of transmission lines. Theoretically, an infinite number of them is required, but practically the situation is not so bad. The examples considered later will illustrate the point. If all the factors of $G(s)$ are also those of $H(s)$, then in equation 18 $b(\alpha)=0$, so that

$$H(s)=G(s)a(\alpha) \qquad (26)$$

Under this special condition the problem is very much simplified. The number of lines required, even theoretically, is finite. It is seen from equation 13 that $(\alpha-1)$ is a factor of both $H(s)$ and $G(s)$.

Case 2

Suppose $T_h<T_g$, or $h<g$. Under this condition, equation 17 can be written:

$$\frac{H(s)}{G(s)}=\frac{p(\alpha)}{q(\alpha)} \qquad (27)$$

where

$$p(\alpha)=p_0+p_1\alpha+\ldots p_{n-1}\alpha^{h-1} \qquad (28)$$

$$q(\alpha)=q_0+q_1\alpha+\ldots q_g\alpha^{g-1} \qquad (29)$$

If α_1', α_2', … α_{g-1}' are the factors of $q(\alpha)$, then equation 27 can be written:

$$\begin{aligned}H(s)=G(s)&(p_0+p_1\alpha+\ldots+p_{h-1}\alpha^{h-1})\\&C'(1-\alpha/\alpha_1')^{-1}\ldots(1-\alpha/\alpha_{g-1}')^{-1}\end{aligned} \qquad (30)$$

In equation 30, the expression

$$C'(1-\alpha/\alpha_1')^{-1}\ldots(1-\alpha/\alpha_{g-1}')^{-1}$$

can be in the form

$$y_0+y_1\alpha+y_2\alpha^2+\ldots$$

so that equation 30 can be written in the form:

$$H(s)=G(s)(z_0+z_1\alpha+z_2\alpha^2+\ldots) \qquad (31)$$

Using equation 16 in equation 31:

$$H(s)=G(s)(z_0+z_1e^{-ST}+z_2e^{-2ST}+\ldots) \qquad (32)$$

Thus $h(t)$ can be realized with a system of transmission lines the input to which is $g(t)$. The remarks made at the end of

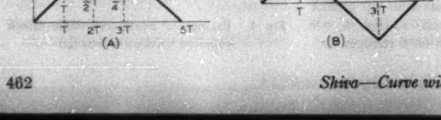


Fig. 3. Generation of h(t) (A) from g(t) (B)

case 1 hold good here also. Two examples may be considered.

Example of case 1. Suppose $h(t)$, shown in Fig. 3(A) is to be generated from $g(t)$, shown in Fig. 3(B). Then, on the basis of equation 8,

$$G(s)=(1/s^2)-(2\alpha/s^2)+(2\alpha^3/s^2)-(\alpha^4/s^2) \qquad (33)$$

and

$$H(s)=(1/s^2)-(1/2\alpha/s^2)-(1/4\alpha^2/s^2)-(9/8\alpha^3/s^2)+(1/8/\alpha^4s^2) \qquad (34)$$

Dividing equation 34 by equation 33,

$$\frac{H(s)}{G(s)}=\frac{-7\alpha-14}{8}+\frac{1}{8}\,\frac{19\alpha^3-16\alpha^2-25\alpha+22}{-\alpha^4+2\alpha^3-2\alpha+1} \qquad (35)$$

In equation 35 set

$$\frac{19\alpha^3-16\alpha^2-25\alpha+22}{-\alpha^4+2\alpha^3-2\alpha+1}=B(s) \qquad (36)$$

Dividing the numerator and denominator of equation 36 by $(\alpha-1)$

$$B(s)=\frac{19\alpha^2+3\alpha-22}{-\alpha^3+\alpha^2+\alpha-1} \qquad (37)$$

Dividing the numerator and denominator of equation 37 by $(\alpha-1)$

$$\begin{aligned}B(s)&=\frac{19\alpha+22}{(1-\alpha^2)}\\&=(19\alpha+22)(1-\alpha^2)^{-1}\\&=(19\alpha+22)(1+\alpha^2+\alpha^4+\alpha^6+\ldots)\\&=22+19\alpha+22\alpha^2+19\alpha^3+22\alpha^4+\ldots\end{aligned} \qquad (38)$$

Using equation 38 in equation 35:

$$\begin{aligned}H(s)&=\frac{G(s)}{8}(-7\alpha-14+22+19\alpha+22\alpha^2+19\alpha^3+\ldots)\\&=G(s)[1+(3/2)\alpha+(11/4)\alpha^2+(19/8)\alpha^3+(11/4)\alpha^4(19/8)\alpha^5+\ldots]\end{aligned} \qquad (39)$$

Using equation 16 in equation 39:

$$H(s)=G(s)[1+(3/2)e^{-sT}+(11/4)e^{-2sT}+(19/8)e^{-3sT}+\ldots] \qquad (40)$$

Equation 40 implies that $h(t)$ can be realized with a system of transmission lines whose input is $g(t)$. It may be pointed out that the coefficients 11/4 and 19/8 alternate throughout the rest of the series. Suppose $h_k(t)$ is defined as the function obtained by considering terms only up to the kth degree term in equation 40, then $h_7(t)$ is as shown in Fig. 4. Comparing this figure with Fig. 3(A), it is seen that there is an unwanted portion in $h_7(t)$. By taking into account more and more terms of equation 40, the unwanted portion can be shifted farther and farther to the right. For instance $h_{14}(t)$ is shown in Fig. 5. Thus, if $h(t)$ is defined by $h(t)=$ a pulse shown in Fig. 3(A) up to $t=5T$ and equal to zero for $5T<t<T_N$, $N>>5$ (and it does not matter what it is beyond T_N), then $h_N(t)=h(t)$. In other words, $h(t)$ can be practically realized for any finite N, or, by having a sufficient number of transmission lines in the system, $h(t)$ can be realized over a working period without any unwanted pulses.

Example of case 2. Suppose $g(t)$ of Fig. 3(B) is to be generated from $h(t)$ of Fig. 3(A). Then, on the basis of equation 8:

$$\frac{G(s)}{H(s)}=\frac{1-2\alpha+2\alpha^3-\alpha^4}{1-(1/2)\alpha-(1/4)\alpha^2-(9/8)\alpha^3+(7/8)\alpha^4} \qquad (41)$$

Dividing the numerator and denominator of equation 41 by $(1-\alpha)^2$,

$$\frac{G(s)}{H(s)}=\frac{1-\alpha^2}{1+(3/2)\alpha+(7/4)\alpha^2+(7/8)\alpha^3} \qquad (42)$$

Instead of factoring the denominator of equation 42, which is bound to bring in numerical error, equation 42 is written as:

$$\begin{aligned}G(s)&=H(s)(1-\alpha^2)[1+(3/2)\alpha+(7/4)\alpha^2+(7/8)\alpha^3]^{-1}\\&=H(s)(1-\alpha^2)\times\begin{bmatrix}1-[(3/2)\alpha+(7/4)\alpha^2+(7/8)\alpha^3]\\+[(3/2)\alpha+(7/4)\alpha^2+(7/8)\alpha^3]^2\\-[(3/2)\alpha+(7/4)\alpha^2+(7/8)\alpha^3]^3\\+\ldots\end{bmatrix}\end{aligned} \qquad (43)$$

Simplifying equation 43:

$$G(s)=H(s)[1-(3/2)\alpha-(1/2)\alpha^2+(5/2)\alpha^3-(25/16)\alpha^4-(51/32)\alpha^5+(47/16)\alpha^6-(1/4)\alpha^7+\ldots] \qquad (44)$$

Introducing equation 16 in equation 43:

$$G(s)=H(s)[1-(3/2)e^{-sT}-(1/2)e^{-2sT}+(5/2)e^{-3sT}+\ldots] \qquad (45)$$

It is evident from equation 44 that $g(t)$ can be generated from $h(t)$ with a system of transmission lines the input to which is $h(t)$. The graph of $g(t)$, using terms up to the seventh-degree term in equation 44, is shown in Fig. 6. By considering more and more terms, just as in the previous example, the unwanted portions can be shifted farther and farther to the right. Taking into account a sufficient number of terms, the unwanted portions can be shifted beyond the working range.

Conclusions

The approximate representation of irregular curves through a series of step functions has been well known but to get a fair approximation, the size of the subinterval has to be small. It may be safely said that in a situation where both the "linear-segment" and "step-function" methods are applicable, the former gives far better approximation with far fewer subintervals in a given total period of time. The linear-segment method, applied in this paper only to functions of finite duration, can be extended to functions of infinite duration also so far as the representation of the function is concerned. Regarding the question of the generation of a function or conversion of one into another, this approach is limited to functions of finite duration only, as it involves time-delay factors. As already pointed out, this method is applicable only if the duration of the function is so small that the corresponding time delays, which are necessary in the generation of the function, can be achieved practically. In other words, the longest time delay which can be obtained practically will be the limiting factor for the application of this method. Laborious algebraic manipulations, for instance the derivation of equation 44 from equation 43, have not been shown as they were not pertinent to the general discussion.

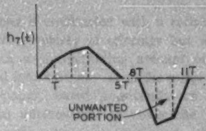


Fig. 4. Illustration of unwanted portion in $h_7(t)$

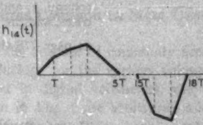


Fig. 5. Illustration of unwanted portion ($h_{14}(t)$) shifted farther to right

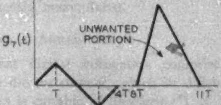


Fig. 6. Graph of $g(t)$ using terms up to seventh-degree term in equation 44

References

1. Synthesis of Passive Networks (book), **Ernst A. Guillemin.** John Wiley & Sons, Inc., New York, N. Y., 1957, chap. 15.

2. Transformation Calculus and Electrical Transients (book), S. Goldman. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1950, chap. 10.

ERRATA

"Universal Modulus—Matrix Logic VI" by E. J. Schubert, published in AIEE Transactions, part I (Communication and Electronics), volume 70, 1960 (January 1961 section), pages 778–81.

The expressions shown enclosed by double parallel bars (page 778, column 3, page 779, column 1, and page 780, columns 2 and 3) are actually binomial coefficients, not matrices, and the bars should be replaced in all cases by parentheses; e.g., *the first equation* on page 778 should read: $_pO_0$: $\binom{p}{0}=1$.

"Tunnel Diode D-C Converter" by H. F. Storm and D. P. Shattuck, published in Communication and Electronics, July 1961, pages 347–52

In Table I on page 351, line 3 in the third column and line 3 in the fourth column should read 5.8 milliwatts and 5 milliwatts (instead of megawatts) respectively.

◆

OUR NEW ADDRESS...

After September 5 we will be located in the new United Engineering Center, at the following address:

AIEE Publications Department
345 East 47th Street
New York 17, N. Y.
(on United Nations Plaza)
Telephone: PLaza 2-6800

A General Description of D-C Digital Voltmeters

CARROLL STANSBURY
FELLOW AIEE

THE WIDESPREAD USE of measurement transducers producing direct-voltage output signals has led to extensive development of analog-to-digital converters, including digital voltmeters. This situation has resulted from the fact that the information output of measurement devices, if handled in reduction processes in other than digital form, progressively deteriorates in accuracy.

Although, in a sense, any analog-to-digital converter is at least potentially a digital voltmeter, the term as used here is applicable to an instrument having as its primary function the direct visual indication of voltage in the form of decimal digits. Accuracy ratings are 1.0% or higher (up to 0.01%). Provision for printer output may be included. Such instruments are essentially direct-voltage instruments, but can be made suitable for alternating voltages by the addition of a converter; it may also be operated as a voltage comparator or a digital ohmmeter by relatively slight modifications of the basic instrument.

The advantages of digital voltmeters include the following:

Feasibility of high accuracy.

Superiority in reading characteristics (no skill required, no parallax, readable from a wide angle or from a distance, no confusion of scales, no interpolation).

Extremely high input impedance (compared with deflection instruments).

Feasibility of automatic polarity indication and range shifting.

Adaptability to code recording (including modifications providing such features as go-no-go indicators, alarms, control systems, and digital data transmission).

Adaptability to modification as an ohmmeter or voltage comparator.

Instead of depending on the relations of mechanical forces or torques as in deflection instruments, the basic mechanism of nearly all digital voltmeters is the comparison of the input voltage with known voltages developed within the instrument. They are thus automatic potentiometers, the higher accuracy designs sharing with manual potentiometers unique advantages in precision voltage measurement, where the accuracy required is higher than can be obtained by deflection instruments, or where the instrument must operate on negligible current.

On the other hand, the digital voltmeter has the disadvantages of high cost, size, and complexity. Its reliability is certainly not yet equal to that of deflection voltmeters. It cannot be sealed, is in many cases a maintenance problem, and involves nonpermanent components. It is inherently more subject to interference from stray fields and from the presence of small alternating voltages superposed on the direct voltage being measured. Many types have one input terminal which in operation must be kept either at power ground potential or can be separated from it to only a limited extent (across low impedance). As presently manufactured, the enclosure usually is conductive and common with the power line ground, in contrast to the insulating molded enclosure frequently provided with deflection voltmeters.

It should be clear, from this brief listing of the advantages and disadvantages of these instruments, that their logical adoption in any given instance will depend on the balance of many factors, including an understanding of operating principles and characteristics. The literature on analog-to-digital converters is extensive, but that on the digital voltmeter itself is not, although such instruments are at present offered by about 40 manufacturers in the United States alone. As a result of the large number of sources of these instruments, the potential user is confronted with a rather bewildering variety of offerings but seems to lack a compensating amount of unbiased information. It is the purpose of this paper, based admittedly on only partial information, to cover the most important operating principles, design features, and performance characteristics of these instruments as a class.

Features Common to Most Designs

In the absence of commonly accepted terminology for the various voltages involved in digital voltmeters, the terms shown in Fig. 1 are arbitrarily adopted in this paper. "Feedback" and "attenuator" as used in this illustration are probably open to criticism and are employed here only because they are frequently encountered.

GENERAL ARRANGEMENT

Most digital voltmeters comprise a basic d-c instrument, which in different makes is capable of measuring 1, 2, 10, or 100 volts directly, connected back of a range attenuator which acts as a voltage divider to give a series of voltage ranges, usually of a maximum value of approximately 1,000 volts. The general arrangement is shown in Fig. 2. Typical ranges are listed in Table I (the letter designations are arbitrarily added for the purpose of this paper only). There are, of course, other ranges available in some models, including low-voltage units in which the maximum voltage is a fraction of 1 volt.

Connected ahead of the range attenuator there may be a low-pass filter, or a converter for a-c operation. In many cases circuits are provided such that the instrument may be used as an ohmmeter of the constant-current type. In others, the instrument is convertible to an automatic Wheatstone bridge.

The input circuit is normally characterized by impedance in the order of 10 megohms on the basic range, with lower impedance when it works at the higher ranges where the limitations of the range attenuator are a factor. Normally, the d-c instrument operates with one input lead tied to an instrument ground which is made common with the power ground, the chassis, and the

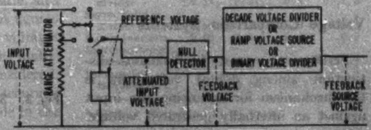


Fig. 1. Suggested designation of voltages

Paper 61-721, recommended by the AIEE Indicating and Integrating Instruments Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 17, 1961; made available for printing April 12, 1961.

CARROLL STANSBURY is with the United States Department of Commerce, National Bureau of Standards, Washington, D. C.

Criticism and suggestions by F. L. Hermach have been of great value in the preparation of this paper.

metallic enclosure. "Floating operation" of the instrument ground is often possible, but usually only at limited voltages between instrument ground and power ground, across a relatively low impedance.

The digital display is either in the form of the familiar columns of neon lights, or as an "in-line" arrangement of numeral indicators of various forms. Where required, the instrument also provides contact closures, corresponding to the digits displayed, for coded signals to operate code printers or for other purposes.

The display also indicates polarity and decimal point position.

Shielding and Grounding

The sensitivity and accuracy involved in many digital voltmeters necessitates close attention to precautions in shielding and grounding. Input leads are shielded, with the shield carried through inside the instrument. Normally, the shield is also connected to the chassis and metallic enclosure. The latter is in turn connected to the power ground via the third connector of the power cord. For floating operation the grounded input lead and the shield may usually be isolated from the chassis and enclosure by an optional connection for measurement at limited voltage difference from ground. Fig. 3 shows a typical arrangement.

Voltage References

Practically all digital voltmeters include a voltage reference source for calibration purposes. Mercury cells have been used to some extent. Unsaturated standard cells have been extensively used and more recently a number of manufacturers have adopted Zener diodes. As compared with standard cells, Zener diodes largely avoid difficulties due to temperature variations and seem to be relatively free from deterioration with age. They appear to be at least comparable in accuracy to unsaturated standard cells.

Specification Items

Information furnished by the manufacturer normally contains a descriptive tabulation or "specification." Appendix I was compiled by collecting items from a number of examples. It includes most of the major descriptive points concerning digital voltmeters.

Voltage Comparison Systems

Perhaps the most fundamental distinction among digital voltmeters is the mechanism for balancing the input against an internally generated voltage. The most commonly used of these mechanisms are briefly described here. The lack of general agreement on terminology complicates the problem of description. In this paper the designations preferred by the author are used, followed (in the headings) by other designations which seem to be commonly used.

Stepping Voltage Divider (Electromechanical, Stepper, Digital Servo)

The filtered, attenuated input voltage is compared with a feedback voltage derived from a decade voltage divider (Fig. 4). The null detector usually involves an electromechanical chopper which actuates stepping relays via an error amplifier. The relays commutate a precision voltage divider (resistor network) supplied by a regulated feedback voltage source. Contacts on the stepping relays provide circuits for operating visual indicators and for external recorders.

The term "tracking" is used to refer to the sequence of logical steps by which the relays attain a position of balance between input and feedback voltages. Tracking takes various forms with different makes and types. The problem is to combine acceptable balancing time with adequate sampling of the input voltage, while avoiding excessive mechanical wear of the relays.

Two types of decade voltage divider circuits for the comparison voltage are common, both borrowed from classical manual potentiometer practice: the Kelvin-Varley circuit and the Wolff circuit. A reference voltage is included either with provision for introducing its output to the normal input circuit of the instrument for calibrating purposes, or with provision for automatic regulation of the feedback-voltage source.

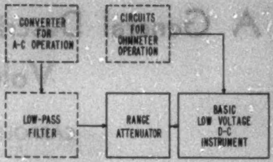

CONVERTER FOR A-C OPERATION
CIRCUITS FOR OHMMETER OPERATION
LOW-PASS FILTER
RANGE ATTENUATOR
BASIC LOW VOLTAGE D-C INSTRUMENT

Fig. 2. Elements of digital voltmeter

Servo Voltage Divider (Analog Servo)

In this type (Fig. 5) the voltage divider which supplies the feedback voltage consists of a motor-driven variable resistor, servo-controlled by a null detector. The resistor brush takes a position such that the input and feedback voltages are balanced. Digital indicators are actuated by mechanical connection to the servo shaft.

Ramp Function (Counting, Sweep, Slope-Time, Time Encoder)

A linearly rising voltage ramp is generated and compared with the input voltage in a differential amplifier (null detector). Coincident with the initiation of the voltage rise, a gate on a clock oscillator is opened and kept open until ramp voltage becomes equal to the input. During this interval, the number of clock pulses that have been passed are counted by an electronic counter circuit. By suitably relating the slope of the ramp voltage to the clock frequency (the time needed for the ramp voltage to increase one voltage unit is made equal to one oscillator cycle

Table I. Typical Voltage Ranges

A	B	C	D	E
0.000–9.999	.0000–.9999	0.0000–9.9999	00.000–99.999	0.0000–1.9999
10.00–99.99	1.000–9.999	10.000–99.999	100.00–999.99	02.000–19.999
100.0–999.9	10.00–99.99	100.00–999.99		020.00–199.99
	100.0–999.9			0200.0–1000.0

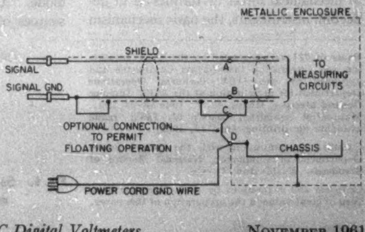


Fig. 3. Shielding and grounding

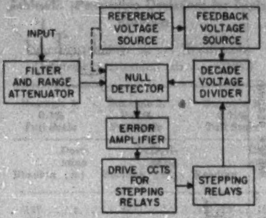


Fig. 4. Stepping voltage divider voltmeter

period) the counter display is made equal numerically to the input voltage (Fig. 6).

In an electromechanical form ("stroboscopic type"), the voltage ramp is generated by continuously rotating the brush of a variable resistor connected across a d-c feedback voltage source. The ramp voltage is balanced against the input voltage; at each instant of equality a gaseous lamp illuminates a numbered drum attached to the rotating shaft, giving visual indication of the voltage.

SUCCESSIVE COMPARISON (SUCCESSIVE APPROXIMATION, SUCCESSIVE TRIALS)

This system makes use of a voltage comparator and a set of precisely regulated voltage sources, one for each binary digit. The source corresponding to the most significant binary digit is first compared with the input voltage: if smaller, it is left in circuit and the next most significant digit compared with the difference between the most significant digit and the input voltage. The process is continued until all of the sources have been tried in decreasing order of magnitude, so that the voltage sources left in circuit at the end of the process represent the input voltage in binary numbers, or in binary-coded decimals. Digital voltmeters operating on this principle are available in all-electronic, mercury-wetted-contact relays, or stepping relay forms.

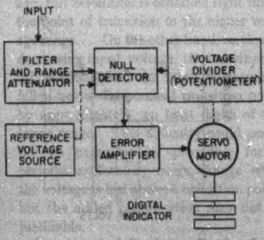


Fig. 5. Servo voltage divider voltmeter

VOLTAGE–FREQUENCY CONVERSION

A voltage-to-frequency converter generates pulses at a rate directly proportional to the input voltage. The rate is determined by counting the pulses during a fixed time interval. In the converter, a charge is accumulated on a capacitor at a rate proportional to the input voltage. Each time the voltage on the capacitor reaches a specific level, a fixed charge of opposite polarity is introduced. The frequency of the latter action (necessary to maintain constant average capacitor voltage) is proportional to the input voltage. Although this is not literally a voltage-comparison system, it is dependent on a precisely regulated voltage as the source of the fixed charge referred to.

CURRENT–TIME CONVERSION

The voltage to be measured is converted to a proportional measuring current. A capacitor is continuously charged by the measuring current, and charged oppositely intemittently by a reference current. The time period within each operating cycle during which the reference current must flow in order to achieve constant average charge is proportional to the measuring current, and is measured by gating and counting a fixed frequency.

CHARGE COUNTING ("ANODIGE")

The input voltage is compared by means of a null detector with a voltage on a capacitor C. C is periodically charged by receiving a number of successive charges from a smaller capacitor, the arrangement being such that the same amount of charge is transferred each time. Thus the voltage on C is proportional to the number of charge transfers (counted electronically). Each time the voltage of C is pumped up to equal the voltage being measured, the null detector stops the process and causes it to repeat itself.

TYPICAL ACCURACY AND SPEED

It is desirable at this point to give some idea of typical values of accuracy and time-per-reading for commercially available instruments, as a function of the type of voltage comparison used. Some data on this point are given in Table II, based on reference 1. As was undoubtedly the case in the original presentation of these figures, they are intended to be only very approximate, and are in no sense to be considered as figures of merit for the corresponding voltage-comparison systems, in view of the multiplicity of other factors to be considered. Furthermore, in the case of the accuracy figures, no attempt has been made in the table to distinguish between percentage reading and percentage full scale, which can be misleading if the figures are taken too literally.

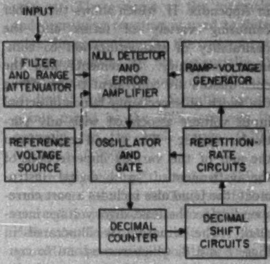


Fig. 6. Ramp function voltmeter

Performance

ACCURACY

There is at present no established standard of terminology used in statements of accuracy by manufacturers, either in the form of such statements or in their exact meaning. However, by analogy with such standards as those of the American Standards Association (ASA) for indicating instruments and for automatic null-balancing instruments (*C39.1* and *C39.4*), the general meaning of stated percentage accuracy can doubtless be taken as "the limit which errors will not exceed" when the instrument is used under "reference conditions."

Accuracy statements regarding digital voltmeters normally are either on a per-cent-of-reading or on a per-cent-of-full-scale basis, depending on the nature of the predominant errors in a particular design, and doubtless on individual opinions regarding suitability. Incidentally, it is surprising how many percentage accuracy statements fail to specify either one.

Typical accuracy statements are listed

Table II. Approximate Accuracy and Speed

System of Voltage Comparison	Accuracy*		Seconds per Reading	
	Best	Usual	Best	Usual
Stepping voltage divider	0.01	0.01	0.2	1.0
Servo voltage divider	0.05	0.2	2.0	3.0
Ramp function	0.05	0.1	0.001	0.1
Switched-binary-reference	0.01	0.01	0.0005	1.9
Voltage-to-frequency		0.2		0.1

* Without distinction as to basis.

in Appendix II, which shows the rather confusing variety of forms and the desirability of standardization to eliminate ambiguity. Statements like the ones quoted are equivalent to specifying the width of uncertainty bands of actual input voltage associated with any particular reading. The spread results from the sum of the various imperfections of the instrument. In a digital instrument, the band also includes a part corresponding to the least-digit voltage increment. The situation is illustrated in Figs. 7 and 8, which correspond to constant and proportional error respectively. The broken line *N* in each figure corresponds to an ideal digital instrument. In order to have such performance, it is necessary that the transition in the digits displayed take place at precisely the input voltage corresponding to points 1, 2, 3, ... Departure from this, such as in the case of line *H* and *L*, may be partly chronic, partly random, in magnitude and sign, and may logically be referred to as analog error.

Analog error may include the summation of such things as the incorrectness of the internal reference voltage, inadequate tie-in of the feedback voltage with the reference voltage, defects in range and in decade voltage dividers, nonlinearity of the voltage rise in ramp-type instruments, relay contact resistance, resistor temperature coefficient, etc.

It has been mentioned that the digitizing process results in an error band one least-digit wide, even in an ideal instrument, entirely apart from the widening resulting from analog error. In an instrument such as the ramp type, where the reading is set up afresh with every reading cycle, this is the only widening of the error band which may be attributed to the digitizing process as such; see Fig. 9(B). By contrast, in voltage-divider-type instruments, the design may be such that once a balance is reached the reading remains unchanged until the input voltage changes from the balance value by approximately one least-digit value; Fig. 9(A). There is no rigid necessity for this overlap to be precisely one least-digit in value: one manufacturer designs for a value of 0.8 least-digit. However, in this type of instrument some overlap is essential to insure stable operation.

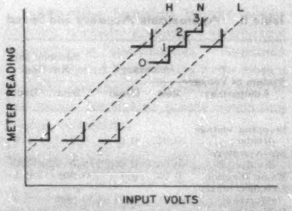


Fig. 7. Meter reading versus input-constant-error basis

Although Appendix II shows the need for standardizing the form of accuracy statement, including simplification if possible, the nature of these instruments seems to make a somewhat complex statement unavoidable. ~~Thus at least the following~~ should usually be included:

1. A percentage figure.
2. Whether per cent full scale or per cent of reading.
3. A statement about "1 digit" or "least count."
4. Range to which applicable.

Items 3 and 4 are omitted by some manufacturers, apparently by using higher percentage figures than would otherwise be used. This has the advantage of simplicity where applicable. The sketch in the lower corner of Fig. 8 shows the special desirability of item 3 in the case of instruments rated on a per-cent-of-reading basis. At low input voltages, this tends to become the predominant part of the error band.

Co-ordination of Resolution with Accuracy, Number of Digits

The readings displayed in digital voltmeters vary both in the number of digits and in the way the different ranges are obtained by changing the location of the decimal point; Table I. The most suitable number of digits in a particular meter depends on the accuracy rating, including whether this rating is on a per-cent-of-reading or a per-cent-of-full-scale basis, and should be such that the resolution is adequate to provide full advantage of the accuracy throughout the range of readings. The provision of more digits than are needed for this requirement may lead to meaningless and perhaps even misleading results.

Tables III and IV are included to show how the above objectives work out in typical cases. The type of accuracy rating is indicated in each case. (In these comparisons, the decimal point can be ignored, since it varies in practice with the range in use, and does not affect the questions considered here.) The lowest reading in each case (such as 1,277) is taken somewhat above 10% of full scale, since below that point there is normally a shift to the next lower range, except when operating on the basic range.

In the case of accuracy ratings on a constant-error basis, most manufacturers use form 1 or 2 of Appendix II; a few use form 4 or 5 in order "to emphasize the digital nature of the readings" or for similar reasons. Standardization recommendations now under consideration call for limiting usage to form 1 or 2. This type of statement is assumed in Table III, which shows imaginary examples 1 and 2 where the accuracy and resolution are consistent. Example 3 shows an inconsistent situation in that the resolution is 10 times the accuracy, so that the last-digit readings are relatively meaningless, although they may be useful when small voltage changes during a short time interval must be observed.

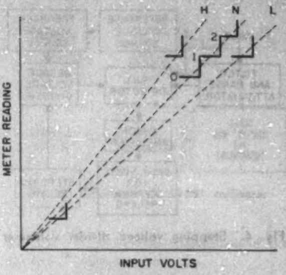


Fig. 8. Meter reading versus input-proportional-error basis

Table IV shows the application of two types of proportional-error accuracy ratings: ±...% of reading, and ±(...%

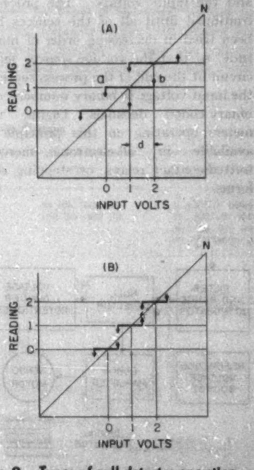


Fig. 9. Types of null detector operation

Table III. Permissible Deviations for Typical Readings

Accuracy rating: ± . . . % full scale.
(Equivalent statement: ± . . . digits.)

Example 1, 0.1% Full Scale		Example 2, 0.01% Full Scale		Example 3, 0.01% Full Scale	
Reading	Deviation (±)	Reading	Deviation (±)	Reading	Deviation (±)
127	1	1,277	1	12,773	10
432	1	4,324	1	43,245	10
997	1	9,976	1	99,763	10

of reading + . . . digit). The table shows for each case the effect of varying the number of digits displayed.

Concerning Table IV the following may be noted:

1. Example *A1* is not suitable for actual use, because it admits no error at the lower readings.
2. Example *A2* shows a satisfactory combination obtained from *A1* by adding another digit in the display.
3. Example *B1* shows a different satisfactory combination obtained from *A1* by adding "±1 digit" in the accuracy statement. *B1* is in effect a compromise between proportional-error and constant-error rating.
4. In examples *A2* and *B2* the true percentage of reading is approximately the same as the nominal percentage given in the accuracy statement. This feature is not essential where the accuracy statement contains a specific constant term, but it seems desirable.
5. Examples *A2* and *B2* differ to a negligible extent in the actual values of permissible deviation.

EFFECT OF VARIATIONS OF INPUT VOLTAGE

An unfortunate effect of the common scheme for obtaining several ranges from a single basic meter may be the existence of a sort of blind spot in the neighborhood of range-transition points. This results from the fact that a shift of range is accompanied by a proportional change in sensitivity. On slowly rising input voltage, full resolution is obtained right up to the point of transition to the higher voltage range. On the other hand, on slowly decreasing input voltage, the transition is not effected until the input voltage has fallen below the nominal transition point by approximately ten least digits of the lower range. This difficulty could possibly be overcome by providing for a periodic exploratory range change when the voltage is just above a transition point, but the added complication may not be justifiable.

In spite of the fact that direct-voltage instruments usually are provided with a low-pass input filter, instability troubles occur at times, particularly if the input voltage includes a frequency which happens to resonate in some way with a natural response frequency of the instrument. Common remedies are the provision of sensitivity adjustment (generally reductions of 10:1 and 100:1) and, alternatively, the use of brief sampling periods instead of continuous sampling. It is claimed for the voltage-to-frequency type of instrument that its inherent integrating characteristic makes it immune to electrical noise and other abrupt changes in input voltage. It will be evident that the sensitivity of most types of digital voltmeters to small superposed components of voltage necessitates care on the part of the user in avoiding applications which lead to unsuspected erroneous results.

A-C VOLTMETERS

Most a-c instruments consist of a basic d-c instrument connected back of a converter in such a manner that the indication of the instrument corresponds to the average or the peak value of the a-c waves. In a few cases, converters are available giving rms output. It is common to derate the accuracy ten to one as compared with the same basic instrument used on direct voltage.

Calibration

The calibration of digital voltmeters is an extensive subject and outside the scope of this paper; therefore only a few remarks will be inserted.

The subject may be divided into the over-all calibration of complete instruments and the components calibration of major parts, such as an input attenuator. Much of the component testing during manufacture is unavoidable, since in the over-all operation each major component must have adequate precision, as there is generally no single adjusting component by which the shortcomings of other elements can be adjusted out in the over-all operation. This situation is no doubt accentuated by the practice of substituting whole subassemblies in digital voltmeters as a maintenance technique.

The usual calibration procedure with deflection voltmeters involves the use of a stable and finely adjustable direct-voltage supply, a precise voltage divider, and a high-grade potentiometer. Tests are made on cardinal points on the lowest range, and on two or three cardinal points on each additional range, all for both increasing and decreasing values of input voltage. (The average of the increasing and decreasing readings is used to determine the corrections for the instrument.) It is convenient to use essentially the same process for the over-all calibration of digital voltmeters. Suitable cardinal points might be the 10% intervals (for example, on the basic range: 0.1000, 0.2000 . . . . 0.9000).

Strictly speaking, the testing of digital voltmeters at a few cardinal points is not as readily justifiable as in the case of deflection instruments. In the latter, provided the cardinal points are not too far apart, it is a reasonable assumption that the corrections to the readings between two cardinal points have a linear relation to the corrections at the cardinal points. On the other hand, in the case of digital voltmeters intermediate readings may involve the action of elements (e.g., resistor steps) not present in the readings at the cardinal points. It follows that a calibration test at selected points leaves an element of chance in the situation, a condition to be avoided with difficulty in an instrument having perhaps 10,000 readings per range, unless by some very elaborate mechanized test setup. In certain cases, for instance some stepping voltage divider or switched-binary reference instruments, it may be possible to obtain the equivalent of testing for all readings by selecting a series of test values in a way that each circuit element is included in one or more tests.

The property of showing all readings of a range in proper order is spoken of as "monotonicity." A test of this property in which all readings are observed would ordinarily be too extensive to be practica-

Table IV. Permissible Deviations for Typical Readings

Comparison of Three Types of Proportional-Error Accuracy Rating

Rating Basis	Typical Reading	Permissible Deviation (±) Digits	Permissible Deviation (±) True % of Reading	Example
±0.1% rdg. 3 digits	127	0	0	A1
	432	0	0	
	540	1	0.2	
	997	1	0.1	
±0.1% rdg. 4 digits	1,277	1	0.1	A2
	4,324	4	0.1	
	5,403	5	0.1	
	9,976	10	0.1	
±(0.1% rdg. +1) 3 digits	127	1	1	B1
	432	1	0.2	
	540	2	0.4	
	997	2	0.2	
±(0.1% rdg. +1) 4 digits	1,277	2	0.2	B2
	4,324	5	0.1	
	5,403	6	0.1	
	9,976	11	0.1	

ble. Here again, however, it is possible to attain an essential equivalent by selecting critical readings. As mentioned previously, digital voltmeters usually include a self-contained reference voltage. In cases where this feature is accompanied by a manual adjustment, the adjustment should undoubtedly be made as a first step in calibration.

GIVEN-READING VERSUS GIVEN-VOLTAGE TESTING

In the normal calibration of deflection instruments, the input voltage is determined for given points on the scale, instead of determining the readings for given input voltages. This is done because the pointer can be more precisely set on a scale mark than it can be read at points intermediate between scale marks.

Although this argument does not apply to digital voltmeters, the given-reading basis is preferable here also. Fig. 9(A) relates continuous (analog) input voltage in the abscissa with discrete readings in the ordinates. If we refer to a given reading "1" we can say it is obtainable for a range of input voltages from *a* to *b* (to a resolution corresponding to that provided in measuring the input voltage, which must necessarily be higher than that of the instrument under test). This is the type of information in which the user is interested, since he starts with a reading and needs to know the band of true input voltages within which it falls.

If on the other hand, we calibrate on a given-voltage basis, such as at some input voltage between 0 and 1, we can only determine the results to the resolution of the instrument under test, by listing the two or more readings obtainable at that input voltage. Also, there may be ambiguity as in the case illustrated, where the number of different readings obtainable for a given input is different at different locations along the abscissa.

Standardization

It has been evident for some time that this field of instrumentation would benefit greatly from a degree of standardization consistent with that existing for the older types of instruments. The most important aspects of needed standardization are in terminology (the need is evident in this paper!) and in the form and exact meaning or accuracy claims. Early in 1960, the Indicating and Integrating Instruments Committee of the AIEE set up a subcommittee for the preparation of a standard. However, interest in the problem by other organizations, particularly the Institute of Radio Engineers (IRE), was so great that the task was promptly passed on to the ASA Sectional Committee on Electrical Measuring Instruments (*C39*), with the active cooperative participation of AIEE, IRE, National Electrical Manufacturers Association, and the National Bureau of Standards. A subcommittee of *C39* has been organized and is now developing a standard. Provision is being made for assistance to the subcommittee by liaison representatives of all of the manufacturers known to be in the field.

Appendix I. Specification Items

1. Input
   Floating?
   Voltage-to-ground limitation
   Impedance
   Filter, a-c rejection
2. Mode of operation
   D-c, a-c, ohms . . .
3. Ranges, range selection
4. Items related to accuracy
   Absolute accuracy
   Relative accuracy
   Stability
   Linearity
   Resolution
   Reference voltage
   Number of digits
5. Operational ambient temperature
6. Response time
7. Presentation (or display)
8. Power requirements
9. Dimensions and weight
10. Mounting
11. Printer output
    Contact closures
    Type coding
12. Supplementary equipment
13. Reading retention
14. Probe
15. Military specification

Appendix II. Typical Accuracy Statements

Constant-Error Basis

1. ±. . .% full scale
2. ±. . .digit (s)
3. ±. . .millivolt (10 mv, etc.)
4. ±. . .% full scale, ±. . .digit
5. ±. . .% full scale ±1/2 last digit ±. . .mv

Proportional-Error Basis

6. ±. . .% of reading
7. ± (. . .% of reading +. . .digit)
8. ± (. . .% of reading or . . . digit, whichever is greater)
9. ±. . .% of reading and/or . . . digit
10. ±. . .% of reading, or . . . mv

References

1. Selection and Application of Digital Voltmeters, **T. E. Nawalinski.** *Electrical Design News*, Englewood, Colo., vol. 5, no. 11, Nov. 1960, pp. S-9 to S-12, S64-5.

2. Accuracy Considerations in Analog-to-Digital Converters for Use in High Speed Systems, **R. E. Wright.** Conference on Analog and Digital Instrumentation, 1959, p. 232.

3. Automatic Printing Ammeter, **T. G. Le Clair.** *AIEE Transactions*, vol. 60, Mar. 1941, pp. 93–98.

4. The Binary Quantizer, **Karl Howard Barney.** *Electrical Engineering*, vol. 68, Nov. 1949, pp. 962–67.

5. High Speed Recording Potentiometer, **R. B. Bowersox, C. G. Hylkema.** *Memorandum no. 20-69*, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, Calif., 1952.

6. Criteria for the Selection of Analog-to-Digital Converters, **G. L. Hollander.** *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 9, 1953, p. 670.

7. A Survey of Analog-to-Digital Converters, **H. E. Burke, Jr.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, no. 10, 1953, p. 109.

8. A Systematic Survey of Coders and Decoders, **B. Lippel.** *Convention Record*, Institute of Radio Engineers, pt. 8, 1953, p. 109.

9. Analog Measurement and Conversion to Digits, **T. C. Fletcher, N. C. Walker.** *ISA Journal*, Pittsburgh, Pa., Sept. 1955, pp. 341–45. (Includes extensive bibliography.)

10. An Unusual Analog-Digital Conversion Method, **B. D. Smith, Jr.** *Transactions*, Institute of Radio Engineers, June 1956, p. 155.

11. A Digital Potentiometer, **S. K. Dean, D. F. Nettel.** *Electronic Engineering*, London, England, Feb. 1956.

12. Analog-to-Digital Converters, **G. G. Bower.** *Control Engineering*, New York, N. Y., Apr. 1957, pp. 107–18. (Includes extensive bibliography and tables.)

13. Definition of Accuracy of Voltage-to-Digital Converters, **B. M. Gorden, I. M. Stone.** *Instruments and Control Systems*, Pontiac, Ill., vol. 32, May 1959, p. 710.

14. An All-Electronic Four Digit Digital Voltmeter, **H. Fuchs, D. Wheable.** *Journal*, British Institution of Radio Engineers, London, England, July 1960, p. 541.

15. A Wide Range Fully Automatic Digital Voltmeter, **J. A. Irving, D. A. Pucknell.** *Ibid.*, p. 536.

16. Digital Voltmeter Logic, **Reginald A. Allen.** *Electronic Equipment Engineering*, White Plains, N. Y., June 1960, pp. 93–97.

17. Digital Voltmeters. *Electromechanical Design*, West Newton, Mass., vol. 4, no. 12, Dec. 1960, pp. 58–75.

Features of an Electronic Crosspoint PABX

J. G. VAN BOSSE
ASSOCIATE MEMBER AIEE

J. DECICCO
ASSOCIATE MEMBER AIEE

THIRTEEN FEATURES characterize the experimental 100-line electronic PABX (private automatic branch exchange), developed at Automatic Electric Laboratories, Inc., Northlake, Ill., a subsidiary of General Telephone & Electronics, Inc. Primarily, the development aims were to gain design and operating experience with electronic switching telephone systems and to test trunking to and from a Strowger step-by-step central office.

System Features

The 13 chief features are:

1. Space-division switching network, using silicon p-n-p-n diodes.[1]
2. One hundred PABX subscribers.
3. Ten 2-way dial trunks.
4. Ten PBX trunks.
5. Six meet-me conference lines.
6. Progressive conference facilities.
7. Four operator circuits on an attendant's cabinet.
8. Trunk-transfer facilities.
9. Trunk restriction for selected subscribers.
10. Night service.
11. Direct trunk facilities, which will allow the operator to link her telephone directly to a trunk, thereby by-passing the electronic exchange in the event of malfunction.
12. Three-digit number scheme in which the first digit is used to access the particular service desired. This PABX uses these digits: 6 for local calls, 8 for conference, 9 for trunks, and 0 for operator service.
13. Telephone modified for tone ringing.

System Description

GENERAL FACTS

A PABX gives telephone service to a business or other organization and also provides facilities for receiving and making calls to a central office. These calls are completed through the same switching unit as local calls. However, the PABX requires an attendant cabinet, for directing incoming calls, and for conference, night-service, transfer, and intercept facilities.

Before describing how to make different types of connections through the PABX, the parts of the system are discussed, and their relationships are shown in Fig. 1.

THE DISTRIBUTOR

Control actions in this system are on a time-division basis. The distributor is a pulse-train generator, supplying a series of pulses to synchronize the different parts. A separate unit, receiving no input from other parts, it can generate independent commands, assuring execution of logical operations in proper sequence.

A wide variety of pulse trains are produced and fed as separate outputs to the rest of the system. Each pulse train has a specific position in time with respect to each other pulse train. In this manner, a certain period of time can be defined by the presence of one or more of the pulses.

The experimental PABX has 130 time slots, each 123 μsec (microseconds) long. The time slots are divided between 100 local subscribers, 20 trunks, 4 operator circuits, and 6 conference circuits. The system cycle is 16 milliseconds, 130 time slots times 123 μsec.

The distributor generates:

1. Tens pulses, equal to 13 time slots or 1.6 milliseconds.
2. Units pulses, equal to 1 time slot or 123 μsec.
3. Interval pulses, equal to 7.7 μsec.

THE MEMORY

This PABX employs a ferrite-core temporary memory of 130 words, one for each PABX subscriber, trunk operator, and conference circuit. Each words has 31 bits, supplying various information about each time slot.

THE MARKER

The marker, a decision-making intermediary between PABX subscribers and the switching network, is a logical network which may be considered stationary with respect to the periodic time reference established by the distributor. It is capable of reviewing the status of each subscriber in turn. Basically, the marker is the means by which information is conveyed from one subscriber time slot to another, and is the logical network that supervises the establishment and termination of audio paths through the switching network.

The marker consists of a number of bistable storage and coding devices (flip-flops), logic building blocks (NOR-gates), a parity checker, and translation matrices. The parity checker is a comparison network that yields an output signal when the input from one source is found to be identical with that from another. Information comes to the switching network in different codings, which require a translation matrix to interpret them and then to transfer the translated information to the switching network.

SUBSCRIBER LOGIC

Subscriber logic[2] is the term applied to a logical network, the function of which is to convert the information in the ferrite-core memory in each subscriber time slot into a form descriptive of his condition. It also acts on this informa-

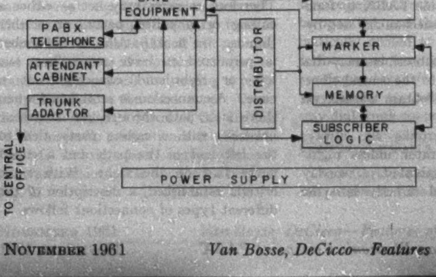


Fig. 1. Block diagram showing relationship of parts of automatic telephone exchange

---

Paper 61-827, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted April 10, 1961; made available for printing April 25, 1961.

J. G. VAN BOSSE and J. DECICCO are with Automatic Electric Laboratories, Inc., Northlake, Ill.

The authors wish to acknowledge their appreciation of helpful suggestions offered by members of the technical staff of Automatic Electric Laboratories, Inc.

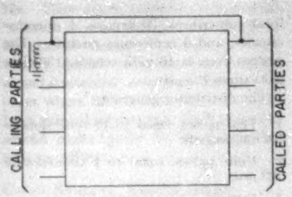


Fig. 2. Switching network

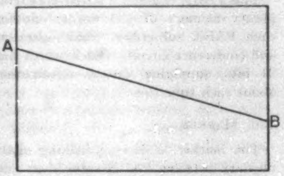


Fig. 3. Local connection

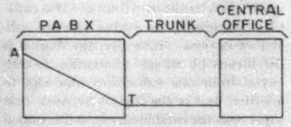


Fig. 4. Outgoing trunk call

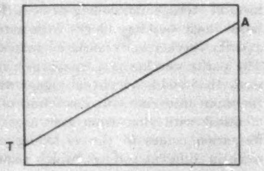


Fig. 5. Inward trunk call

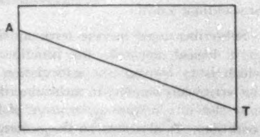


Fig. 6. Inward trunk connection reversed

tion, as well as on information from the marker and multiplex highway, in such a manner as to determine what information should be written into the memory at the end of each time slot. It is used by each subscriber during the time assigned him.

Subscriber logic is composed of bistable storage elements (flip-flops) and logic building blocks (NOR-gates).

Switching Network

Providing a means of establishing and maintaining an audio connection between selected lines and trunks, the switching network consists of a 4-stage crosspoint array, which uses p-n-p-n Shockley diodes as crosspoint switches. Each stage consists of groups of matrices made up of vertical and horizontal rows of connections with a p-n-p-n diode connected at every crosspoint. Each subscriber's line equipment is connected to both ends of the network, permitting him to act as a calling or called party, or both. A split junctor is used in the center of the network for applying part of the potentials to the crosspoints, gating the ringback and ringing signals, and supplying the necessary holding current to the p-n-p-n diodes. In addition, the switching network provides facilities for breaking down the audio connection, once subscribers have concluded their conversation.

Line Equipment

Each PABX subscriber, PBX trunk, 2-way dial trunk, operator circuit, and conference circuit has an associated line circuit. The line circuit terminates the line loop and also provides an audio connection to the crosspoint switching network. The line circuit has sensing elements that reflect the conditions of the line loop; by opening and closing the line loop with either the hookswitch or dial springs, a subscriber may make his intentions known. The line circuit will also determine when certain supervisory signals should be supplied to a subscriber.

Trunk Circuits

There are facilities for ten 2-way dial trunks, permitting in-and-out-dialing, and ten PBX trunks, on which all incoming calls are directed to the attendant cabinet. Associated with the trunks are a number of NOR-gates, flip-flops, relay drivers, and relays. These components and circuits are arranged in such a way as to enable the electronic PABX to function with the electromechanical step-by-step central office.

Only unrestricted subscribers may dial directly over trunks to the central office; restricted subscribers obtain a trunk via the operator.

Incoming PBX trunks are always directed to the operator unless night-service facilities are activated. Two-way dial trunks may dial directly into the PABX.

Attendant Cabinet

In this system, the attendant cabinet is a cordless, desk-mounted turret-type cabinet. Besides providing regular attendant service on incoming calls, it provides night-service, conference, transfer, and intercept facilities.

There are four operator positions, each comprising a conference key, a 3-position answer key (NORMAL, ANSWER, and HOLD), and two lamps (local and trunk) which flash when operator action is desired and when attendant answers.

Power Supply

The transistorized power supply was designed to provide a number of regulated voltages with suitable means of monitoring, sequencing, and protecting the system.

The system draws its power from the a-c mains. A line-voltage regulator eliminates the voltage fluctuations normally encountered on the mains; hence, the series voltage regulators can be made to operate more marginally with respect to input voltage while higher efficiency of the supply is maintained. The regulated d-c voltages are of two categories, critical and noncritical. Both types are applied to the exchange through power switching devices. The START-STOP operation of the control equipment is sequential, and insures that noncritical voltages are never present on the exchange buses in the absence of critical voltages. A power-monitoring panel is used to determine the accuracy of all voltages. A voltage-sensing circuit provides automatic protection against excessive deviations in critical voltages.

Establishing Connections

General Criteria

The means for establishing the various connections can be shown by depicting the network in a simplified form, as in Fig. 2. All the line equipment associated with PABX subscribers, operator positions, and trunks is physically tied to both sides of the switching network. This is shown in Fig. 2 for one of the line circuits. Therefore, a caller may act as either a calling or a called party, or neither (handset on hook). Also, no subscriber is permitted to have more than one left- or right-hand connection at a time. A connection is established when there is a path through the switching network, with a calling party tied to the left end of the path and a called party tied on the right. With these criteria established, a description of the different types of connections follows.

LOCAL CONNECTION

Let us assume that local subscriber *A* (calling party) wishes to call another local subscriber *B* (called party). When *A* removes his handset from the cradle, the common-control subscriber logic will review his status and, if conditions are proper, he will order the dial tone returned to him, directing him to commence dialing.

By dialing the access digit 6, a PABX subscriber makes known to the subscriber logic that he wishes to make a local call. Upon dialing the second and third digits, which represent the "tens" and "units" of of the called party's address, a demand for marker action is established. If the subscriber meets all conditions as an eligible party, the marker will locate, route, and establish a connection. The calling party will be connected on the left and the called party will be connected on the right-hand side of the switching network, as in Fig. 3.

Ringback tone and ringing frequency will be returned to the calling party and called party, respectively. After *B* answers, he will have established an audio connection through the crosspoint network to calling party *A*.

OUTGOING TRUNK CALL

Let us assume that local subscriber *A* (calling party) wishes to make a call over trunk *T* (called party). If *A* has a restricted status and attempts to be connected to a trunk by dialing the access digit 9, he will be routed to the operator. If the calling party is unrestricted and he is not already connected on the left as a calling party, he may dial the access digit to request a trunk. The subscriber logic will then direct the marker to locate the first available trunk (it can be either a PBX or a 2-way dial trunk) and establish a connection.

After the connection is made by the marker, the trunk loop will be closed, causing the MAX (distant central office) to be seized. Dial tone will be returned to the calling party. Also, a tone gate in the calling party's line circuit will be energized to permit a high-frequency tone, under the control of his dial, to pass through the transmission network. This interrupted high-frequency tone is detected in the selected trunk's line circuit and converted to d-c pulses (similar to standard dial pulses) for pulsing a relay to open and close the trunk loop. Upon dialing the 7-digit address of the distant MAX, an audio connection is established. The left- and right-hand connections and the transmission path are shown here in Fig. 4.

INWARD TRUNK CALL AND TRANSFER

Let us assume an incoming call from a 2-way dial trunk *T* (calling party) to a PABX subscriber *A* (called party). Also, let us assume that, after the connection is made, the calling trunk *T* requests a transfer to subscriber *B* in the PABX.

Since these trunks are accessed by the PABX via the fifth selector in the Strowger step-by-step central office, only the last two digits will be pulsed into the relay trunk adapter. There they will be recognized by the PABX as the "tens" and "units" of the desired called subscriber. If the called subscriber *A* meets the required conditions, he will be connected as shown in Fig. 5.

If the calling trunk wishes to be transferred to PABX subscriber *B*, a problem becomes evident. Since the trunk *T* is already a calling party, he may not again act in that capacity. The PABX will recognize this dilemma and create another connection where *A* will act as the calling part and *T* as the called party. This is shown in Fig. 6.

The trunk may now act as a calling party since it appears as a called party, although *A* will physically dial *B*'s number. Upon completion of dialing *B*'s number, the new connection will be as shown in Fig. 7. This condition is acceptable to the PABX, because trunk *T* has only one connection as a calling party and one connection as a called party.

When the connection is made as shown in Fig. 7, the trunk *T* is silenced. If subscriber *B* does not answer, subscriber *A* can dial the digit 2 (which will "un-silence" *T*) and notify him of *B*'s condition. If subscriber *B* does answer *A*, and agrees to talk to *T*, then *A* can simply remove his connection by hanging up. This action by *A* will destroy *A*'s connections to *T* and also un-silence *T*; the connection then will be as shown in Fig. 8.

If the incoming trunk call had been a PBX trunk, the actions would be similar, except that incoming PBX trunks are seized by the attendant. The transfer operations would be the same as performed by called party *A*. If the incoming PBX call had been received in the absence of the attendant, then the night-service key on the attendant cabinet would be operated, and the call would be automatically routed to a local PABX subscriber wired for night service. The answering and transfer action would then be very similar to those of the incoming 2-way dial trunk connection.

CONFERENCE CALLS

There are six conference lines; they are passive and are tied to only the right-

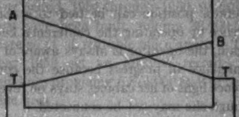


Fig. 7. Inward trunk transfer established

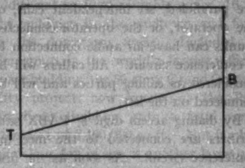


Fig. 8. Inward trunk transferred

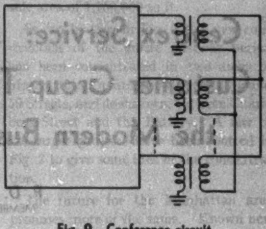


Fig. 9. Conference circuit

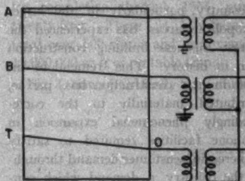


Fig. 10. Conference call

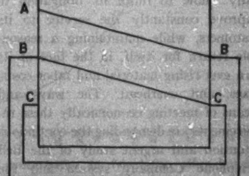


Fig. 11. Progressive conference call

hand side of the ~~network~~. Also, the operator position can be tied to a conference by operating the conference key. Again, the operator is always aware of a conference in progress because the conference light of her cabinet stays on, even though she may not be connected.

The conference lines are connected as shown in Fig. 9. They can act only as called parties, since they are tied only on the right of the network. Since the secondaries of the conference transformers are paralleled, six independent calls, or the operator, or the operator-connected trunks can have an audio connection to a conference circuit. All callers will be considered as calling parties and will be connected on the left.

By dialing access digit 8, PABX subscribers are connected to the meet-me conference circuit. As soon as the first conference circuit is connected, the conference light on the attendant's cabinet will light, signaling that the circuit is in use. The connections on the left and the right of the switching network are as shown in Fig. 10.

Progressive Conference or Chain-Call Connection

If PABX subscriber *A* calls PABX subscriber *B*, and in turn calls *C*, etc., this type of connection ties a number of subscribers together progressively.

PABX subscriber *A* calls *B*, as explained previously, in a local call. Since *B* is connected only on the right, as a called party, he is eligible as a calling party if he desires. Subscriber *B* calls subscriber *C* by dialing *C*'s number. Subscriber *B* is now connected as a called and calling party—that is, connected on both the right and left. Subscriber *C* can call subscriber *D*, etc.

These types of chain calls can be broken in the middle of the chain and have two independent audioconnections if there are enough subscribers in the chain. For example, in a chain involving *A*, *B*, *C*, *D*, and *E*, subscriber *C* can hang up and a connection will still exist between *A* and *B*, and *D* and *E*. If *B* had hung up, *A* would also have to, but there still would be a chain connection between *C*, *D*, and *E*. This is illustrated in Fig. 11.

References

1. Transmission Network of an Electronic Crosspoint PABX, R. F. Kowalik. *AIEE Transactions* (see pages 491–96 of this issue).

2. Logical Control of an Electronic Crosspoint PABX, R. P. Sanders. *Ibid.* (see pages 496–501 of this issue).

Centrex Service: A New Design for Customer Group Telephone Service in the Modern Business Community

P. D. SHEA
MEMBER AIEE

SINCE the end of World War II, this country, particularly in the larger metropolitan areas, has experienced the greatest business building construction boom in history. This tremendous increase in new construction has, per se, contributed materially to the correspondingly phenomenal expansion in telephone facilities required to satisfy the increasing customer demand throughout the country, a demand which continually calls for newer, better, and speedier telephone service at prices within reasonable reach of existing levels.

Obviously, such a situation presents a serious problem to the telephone company: how to fulfill its obligation to improve constantly the service to its customers, while maintaining a reasonable return for itself, in the face of its own ever-rising material and labor costs, taxes, and overhead. The ways and means of meeting economically these requirements are demanding the continuous attention and urgent study of the Bell Telephone Company system and the entire telephone industry.

Because of the magnitude of the overall telephone requirements, as they are influenced by the building boom, this paper will consider the problem only as it is related to:

1. The service area in which the effect of new building has probably been felt most acutely, namely, the Private Branch Exchange (PBX).
2. The geographical area of Manhattan Island, New York City, where the problem has been magnified by the great number of new commercial office buildings constructed in the last decade. It should be pointed out, however, that the same basic problem exists in many of the other metropolitan areas throughout the country. Only in magnitude and volume of construction is Manhattan different.

New and imaginative designs have begun to make definite progress toward alleviating the problem in the PBX area. A recently developed concept for serving all PBX customer groups at any given location on a centralized basis will replace, in many instances, the conventional, individual type of PBX service. Christened Centrex Service by the Bell System, this new concept includes several variations in the proposed methods of service. For the purpose of this paper, however, the discussion will be limited mainly to two specific methods of serving groups on a centralized basis:

1. Using the no. 5 crossbar switching system (redesigned) as a central office to provide direct-line service to the stations of customer groups in one or more buildings which would normally be served by step-by-step PBX's on the customer's premises. The no. 5 crossbar unit can be located in a nearby telephone company central-office building, in rented quarters in a commercial office building, at an airport, etc.
2. Using a centralized *701B* step-by-step PBX system located in telephone company space, usually in rented quarters in a commercial office building, to serve all the customer groups in that building as well as those in adjacent buildings.

The discussion material will be presented in the following sequence: First, in order to understand the full impact which new building construction has on PBX requirements, a brief picture will be given of the situation in the Manhattan area covering the post-war period, with a short look into the 1960's. Next, a rather broad technical discussion of the no. 5 crossbar and *701B* centralized designs will be presented, followed by the practical application of these new concepts to the telephone industry, with a brief story of the impact which customer group centralization has had on the Bell System. Finally, the conclusions will be stated.

Paper 61-722, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 22, 1961; made available for printing May 3, 1961.

P. D. Shea is with the New York Telephone Company, New York, N. Y.

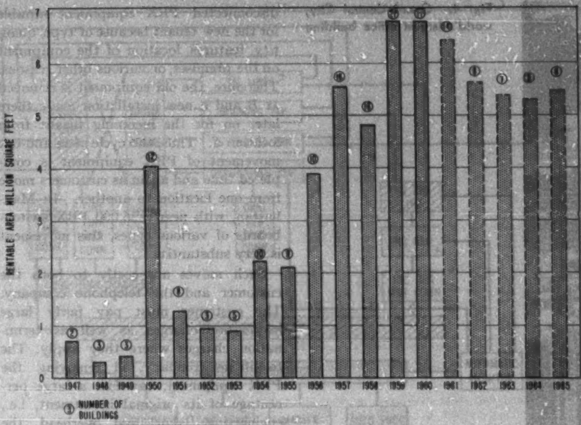


Fig. 1. Post World War II Manhattan office building construction

Cause and Effect

CAUSE

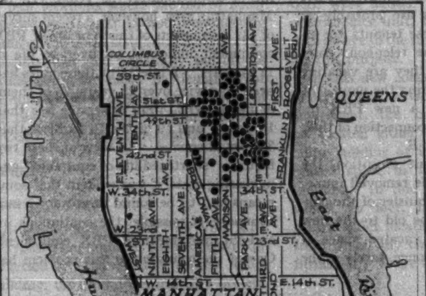
Since the end of World War II, the amount of new commercial office space constructed in Manhattan has been enormous. This has been particularly true during the last 7 years when the pace has been stepped up considerably, as can be seen in Fig. 1. For example, in the previous 7 years, 1947–53, 36 new office buildings were completed with a total rentable area of nearly 9,000,000 sq ft (square feet). However, in the last 7-year period, 1954 through 1960, 93 buildings have been built, adding 32,000,000 more sq ft of rentable office space. The year 1959 set an all-time record: 21 buildings providing 6,874,000 sq ft of rentable area.

As tremendous as these figures are for the 14-year period, the pace has slackened very little since. Already under construction for completion in 1961–62 are 20 more office buildings with 12,000,000 additional sq ft of rentable space, 16 of the buildings in the midtown and four in the downtown areas. Included in the group is the world's largest office-space building, the 59-story Grand Central City project, now known as the Pan American Building, with a total rentable area of 2,400,000 sq ft. Not far behind Pan Am in size is another large one at 1290 Avenue of the Americas, a 43-story structure which has been named the Sperry Rand Building, with a total rentable area of 1,700,000 sq ft.

As can be seen in Fig. 2, the new construction of the entire postwar period has been concentrated in two areas of Manhattan: midtown, between 34th and 59 Streets, and downtown, between Chambers Street and the Battery. A few of the larger projects have been spotted in Fig. 2 to give some idea of the concentration.

The future for the Manhattan area promises more of the same. Known new office building projects on the drawing boards or in the very early planning stage for completion through 1965, number twenty-four. These 24 buildings will total over 500 stores and approximately 15,000,000 sq ft of rentable area.

Fig. 3 shows graphically the cumulative


COLUMBUS CIRCLE
59th ST.
51st ST.
49th ST.
42nd ST.
W. 34th ST.
34th ST.
23rd ST.
W. 14th ST.
E. 14th ST.
ELEVENTH AVE.
TENTH AVE.
NINTH AVE.
EIGHTH AVE.
SEVENTH AVE.
AVE. OF THE AMERICAS
FIFTH AVE.
MADISON AVE.
PARK AVE.
LEXINGTON AVE.
THIRD AVE.
SECOND AVE.
FIRST AVE.
FRANKLIN D. ROOSEVELT DRIVE
BROADWAY
MANHATTAN
QUEENS
East River
Hudson River
NEW JERSEY
HOUSTON ST.
DELANCEY ST.
GRAND ST.
CANAL ST.
WORTH ST.
CITY HALL
FULTON ST.
PEARL ST.
WALL ST.
LAFAYETTE ST.
BKLYN.
0 MI. 1

New York Times

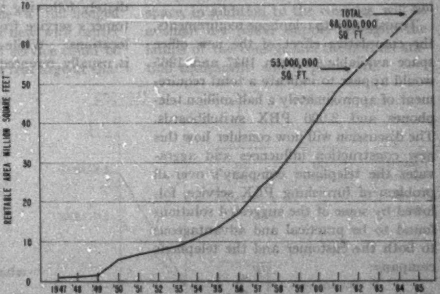


Fig. 2 (left). Areas of concentration for Manhattan office building construction

Fig. 3 (above). Manhattan office building cumulative construction

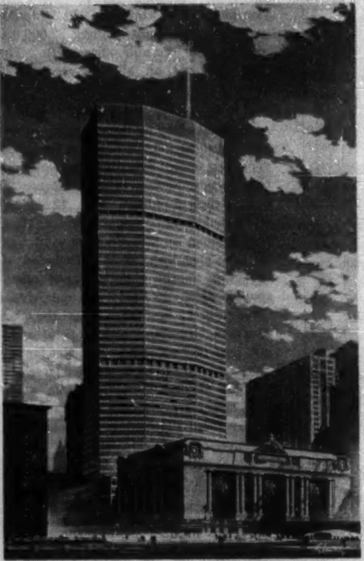


Fig. 4. Grand Central City, world's largest office building

effect of 14 years of office-space construction. This gives a rather vivid picture of the cause of a serious problem to the New York Telephone Company. Similar problems have been present for other telephone companies throughout the country.

In the period 1947 through 1960, these new buildings have competed in the open rental market for tenants; the few exceptions being buildings erected by corporations for their own use. In almost every instance, all space has been fully rented before the completion of the building.

Translated into telephone requirements, the cumulative effect of the new office space available between 1947 and 1965 would appear to indicate a total requirement of approximately a half-million telephones and 2,000 PBX switchboards. The discussion will now consider how this new construction influences and aggravates the telephone company's over-all problem of furnishing PBX service, followed by some of the suggested solutions found to be practical and advantageous to both the customer and the telephone company.

EFFECT

Obviously, as a newly constructed office building nears completion, new tenants plan to move in to occupy the office space (location *A*). The effect of this move, almost without exception, is to start a chain reaction of in-and-out movement of PBX equipment. The odds are probably at least 100 to 1 that these tenants are already customers of the telephone company at the premises they are vacating (location *B*). In planning the move, the customer has ordered a new PBX installed at *A* and the disconnection of their present PBX service at *B*.

Again, almost without exception, the PBX equipment at *B* is removed immediately following the transfer of the customer's service from the old to the new locations. While the vacated space *B* is usually rerented at once, rarely is the disconnected PBX equipment suitable for the new tenant because of type, quantity, features, location of the equipment on the premises, or various other reasons. Therefore, the old equipment is removed at *B* and a new installation made there later on for the incoming tenant from location *C*. Thus, the cycle of in-and-out movement of PBX equipment is completed time and again as customers move from one location to another. In Manhattan, with nearly 25,000 PBX switchboards of various types, this movement is very substantial.

Such moves are costly to both the customer and the telephone company. The customer must pay fairly large installation charges, as well as termination charges where they apply. The telephone company, in removing the PBX equipment, can lose a large percentage of its original investment, i.e., engineering, labor, and overhead, the extent of its loss depending, of course, upon the length of time the equipment has been in service. This loss has at times run as high as 60% of its original investment.

This continuous movement of PBX equipment in Manhattan, particularly in the last 10 years, has had a very definite effect on life. For example, the *701* type *S×S* dial PBX equipment of the New York Telephone Company has an average life of approximately 22 years. In the Manhattan area, however, a recent survey of 132 disconnects of this type of *701* PBX equipment indicated an average life of 11$^1/_2$ years. While 14 of these units had been in service over 25 years, 76 of them, or nearly 60%, were in service less than 10 years; and 19 of the 76 had less than 3 years of service. In view of these two factors: shortened life of equipment and increasing costs, it was evident that some means had to be found to cut down and eliminate, if possible, the continuous in-and-out movement of PBX equipment. One of the solutions to this problem has

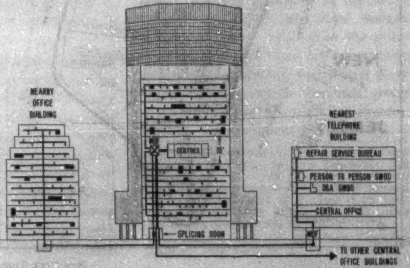


Fig. 5. Typical Centrex arrangement in rented space

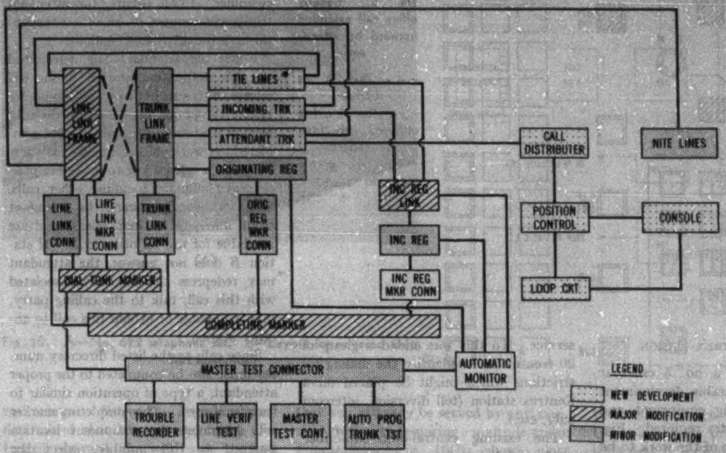

TIE LINES*
INCOMING TRK
ATTENDANT TRK
ORIGINATING REG
LINE LINK FRAME
TRUNK LINK FRAME
LINE LINK CONN
LINE LINK MKR CONN
TRUNK LINK CONN
ORIG REG MKR CONN
INC REG LINK
INC REG
INC REG MKR CONN
CALL DISTRIBUTER
NITE LINES
POSITION CONTROL
CONSOLE
LOOP CKT
DIAL TONE MARKER
COMPLETING MARKER
MASTER TEST CONNECTOR
AUTOMATIC MONITOR
TROUBLE RECORDER
LINE VERIF TEST
MASTER TEST CONT.
AUTO PROG TRUNK TST
LEGEND
NEW DEVELOPMENT
MAJOR MODIFICATION
MINOR MODIFICATION

Fig. 6. Block diagram of new and modified circuits required to provide Centrex service on no. 5 crossbar

* Various types of lines grouped in one block to simplify

been the development of centralized customer group service, or Centrex, which will gradually replace the conventional individual types of PBX service. As previously mentioned, two concepts of Centrex service have been designed for this purpose: (1) the no. 5 crossbar design for the larger installations, and (2) the centralized *701B* step-by-step design for new locations where the use of no. 5 crossbar equipment is not economically sound or where the structural design of the building does not physically permit its installation; or, more particularly, for the conversion of existing installations of *701* type step-by-step PBX installations to Centrex wherever it is practicable.

Solution

Centralization

About 2 years ago, the New York City newspapers announced the proposed construction of the world's largest office building, the Grand Central City project, Fig. 4. This news release was of very great interest to the Bell System engineering people who had been devoting much time and effort toward advancing a new design aimed at eliminating costly movement of the conventional types of PBX equipment.

In this new approach the customers located in large office buildings were regarded as telephone subscribers located in a civic community, and to be served accordingly, i.e., on an individual line basis from central office-type equipment, modified, of course, to provide the features normally furnished with PBX service, such as 4-digit extension-to-extension dialing ("intercom"), transfer to another station of the same PBX customer, etc. Also, new features such as Direct Inward Dialing (DID) to the PBX station, new pushbutton attendant position consoles with illuminated keys and arranged for keypulsing completion of in-and-out calls, and full automatic number identification, were added. The no. 5 crossbar central-office unit, having been designed with an eye to the future, seemed a natural for this purpose.

In this new concept of providing customer group service from central-office equipment, the equipment itself can be located either in a nearby central-office building or in telephone company rented space in the commercial office building. In this last case, the location would, in fact, become a full-fledged central-office location, a new wire center from which any class of service could be furnished. There was a real urge at that time in the Bell System to make a Centrex installation in rented space and the Grand Central City project certainly appeared to be an ideal candidate. Consequently, steps were taken to implement the development work necessary to arrange the standard no. 5 crossbar equipment so that it could be used for both centralized group service and regular central-office service.

Fig. 5 depicts the centralized customer group service in a commercial and nearby office buildings with the Centrex unit located in rented space. Fig 5 also shows the Centrex relationship to the normal central-office-type services, such as repair bureau, traffic operating switchboards, etc. The attendant's position equipment will be located on the customer's premises. In fact, the only telephone equipment to be located in the customer's space, in addition to the consoles, will be a small a-c power supply to light the lamps of the console and the normal station equipment, wiring plans, etc.

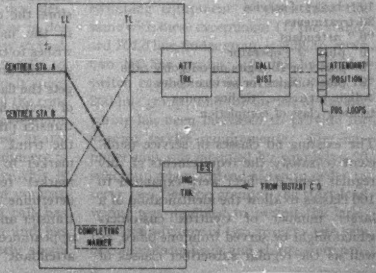

LL
TL
CENTREX STA A
CENTREX STA B
ATT TRK
CALL DIST
ATTENDANT POSITION
POS LOOPS
ICS
INC TRK
FROM DISTANT C O
COMPLETING MARKER

Fig. 7. Centrex office transfer-type call

— — — Incoming calls
– – – To transferred station
–·–·– Incoming trunk to attendant

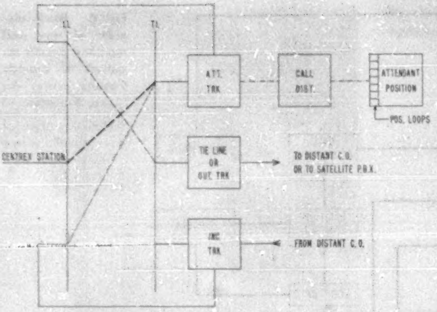


Fig. 8. Centrex office call extended outward by attendant

- - - Dial "0" or transfer
— — — Extended outward call from attendant
—·—·— Incoming trunk to attendant

No. 5 Crossbar Centrex Design

Centrex service in a no. 5 crossbar office required considerable development to modify the existing equipment and provide the new circuitry required. Fig. 6 shows the magnitude of the work to be done. Since the basic idea was to provide a modified individual line-type of service for Centrex stations, each of these stations was given a line link-frame appearance. This allowed direct station completion and automatic station billing without any modification of existing equipment. Other features, however, such as transfer, attendant operation, intercom service, and tie-line completion, normally handled by standard PBX equipment, had to be added to the regular central-office operations without affecting the service provided for non-Centrex subscribers.

Certain basic changes in the common equipment were required to accomplish this. The service features provided are as follows: for standard no. 5 crossbar central office units:

60 classes of service
"0", operator
7- or 10-digit outgoing calls
3-digit service codes
no transfer

For Centrex central office units:

100 classes of service
20 treatments
"0," attendant
"9" plus "0", operator
"9" plus 7 or 10 digits on outgoing calls
"9" plus 3 digits for service codes
4-digit intercom; tie-line codes
transfer class of completion

The existing 60 classes of service (sufficient to satisfy the requirements of any regular central office) were expanded to 100 classes to allow the identification of a larger number of Centrex customers which might be served from one office, as well as the regular subscriber classes of service. To this was added a group of 20 treatments to identify the dialing restrictions which might be placed on a Centrex station (toll diversion, intercom only, etc.).

The existing central-office-subscriber method of dialing 0 to reach the operator, 7 or 10 digits on outgoing calls, and 3 digits for service codes, had to remain unchanged for regular subscribers. Yet, when the Centrex station dials 0, the call must be completed to the attendant; for outgoing calls to a central office, the prefix 9 is dialed; and each station has access to 4-digit intercom and tie-line codes. To determine the proper destination on a call, the code-translation field was modified: first, to screen through the class of service to determine the type of customer; second, if it is a Centrex station it screens through the treatment field to indicate whether or not the station is permitted to dial this particular code; and third, it selects the proper route for the call.

Another major modification of the common equipment was the provision for a transfer type of call. This type of operation is best explained by tracing the path of such a call; see Fig. 7.

An incoming call to any station is completed in the normal manner. However, if the destination is a Centrex station, station *A*, the completing marker will store the class of service of the station in the incoming trunk. If station *A* wishes to transfer the call, it will flash the switchboard. When the incoming trunk gets the flash, it first checks to see if class of service has been stored (indicating that transfer privileges are allowed) and, if so, the trunk then calls in the completing marker to set up the connection. The marker "reads out" the class of service to determine the customer requesting the transfer and connects the line link-frame appearance of the incoming trunk to an attendant trunk associated with that customer. Upon seizure, the attendant trunk will connect to an idle attendant through the distributing circuit.

When the attendant has determined the new destination, station *B*, she will release the connection to station *A* and, by keying the proper digits, set up the connection to station *B*. Since she has completed her work on this call, she may now depress the release key to make her position available for handling other calls. When station *B* answers, the attendant trunk and position loop release and are available for use on other calls. If station *B* does not answer, the attendant may redepress the talk key associated with this call, talk to the calling party, take a message or extend this call to another station.

Since calls to the listed directory number must also be completed to the proper attendant, a type of operation similar to transfer is used. The completing marker will determine the equipment location assigned to this number, select the proper line-link appearance to determine the associated class of service, and connect the line link-frame appearance of the incoming trunk to an attendant trunk associated with this customer. The principal difference is in the signal given the attendant trunk which notifies the attendant that this is a listed directory call.

There are times, however, when the call will not be completed locally, such as attendant-assisted outgoing calls and incoming calls completed to a tie line; Fig. 8. In such cases, the attendant trunk will be selected by a station dialing 0 or a transfer-type call. The trunk will then connect to the attendant in the nor-

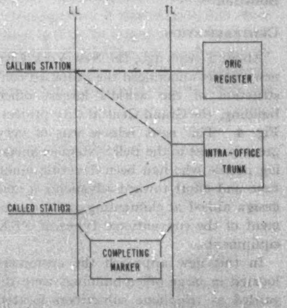


Fig. 9. Centrex office intercommunication call

- - - Dial pulses
— — — Station to station of same customer

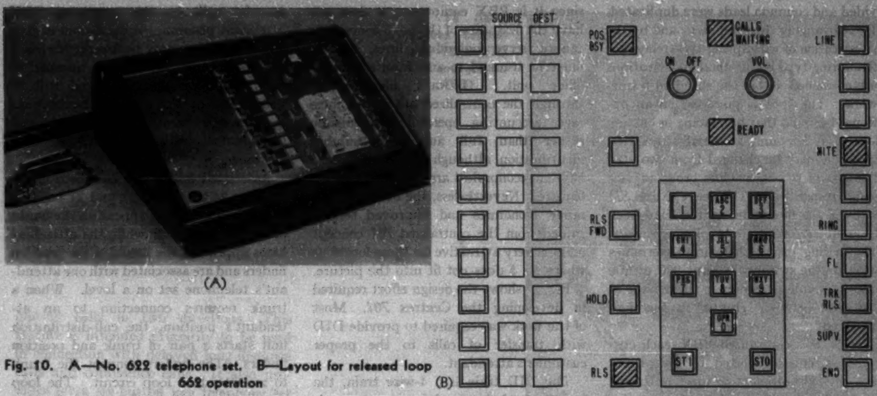


Fig. 10. A—No. 622 telephone set. B—Layout for released loop 662 operation

mal manner. When the attendant has determined that the call should be extended outward, she will cause the attendant trunk to use its line link-frame appearance. In this way, the attendant may extend the call to any equipment available by dial selection and, although the position and position loop are released to serve other calls, the attendant trunk is held for the duration of the call.

The remaining change in basic operation is the 4-digit intercom call; Fig. 9. To restrict this type of dialing to calls between two stations of the same Centrex customer, the originating register is informed of the class of service of the calling station when it is attached to receive the dial pulses. At the end of dialing, the originating register will pass the class information to the completing marker. If this is an intercom-type call, the marker will compare the classes of service of the calling and called station and the call will be allowed to complete only if they match, indicating two stations of the same customer.

Completely new attendant equipment has been designed to keep pace with the Centrex concept of operation. The attendant console, as shown in Fig. 10(A), has been designed as an interim facility, as far as its physical appearance is concerned, pending the availability of the system's new universal console design, which can be expected shortly. However, the basic circuit design and operating features of this console have been incorporated in the new universal version. The console as shown in Fig. 10(A) is now available for either Centrex or centralized *701* dial PBX service.

Incoming calls are connected to an idle console through distribution equipment so that any call may be served by any position without requiring multiple-trunk appearances. Fig. 10(B) shows more clearly the face details of the console. At the left are the ten loop keys with their associated SOURCE and DESTINATION lamps to indicate station supervision. These keys are used to terminate incoming calls to the attendant on a "released-loop" basis, which means that on all calls of this nature the loop will be released from the connection upon satisfactory completion of the call by the attendant. Only four to six loops are normally equipped since this amount of traffic is as much as one attendant can handle. The remaining keys may be used to terminate tie lines or other special equipment.

A row of common keys is shown second from the left. The RELEASE key is depressed to indicate that the attendant has completed her work on a call and to make the position available to receive other calls. The HOLD key is used to hold a call on the console if the attendant wishes to re-enter the connection after having released it. The RELEASE FORWARD key is used to release an existing connection so that the attendant may connect the calling station to a new destination. The POSITION BUSY key is depressed to make the position busy so that the distributer will not connect calls to this position while the attendant is busy on other work. At the center of the console is the key set used by the attendant for pulsing with its associated READY lamp which lights when the equipment is ready to receive pulses. Above this is the AUDIBLE SIGNAL OFF–ON and VOLUME control which indicates that a call has been connected to the position and has not been answered. Immediately above these is a position CALLS WAITING lamp. Other lamps are provided for a supervisor's desk to indicate roughly the number of calls waiting.

On the right is a strip of ten keys which are used for miscellaneous signaling functions. The spare keys may be used on an optional basis to provide other features such as conference circuits or night connections.

701B Step-by-Step Centrex Design

Although great impetus was placed behind the no. 5 version, and rightly so, it was realized that a step-by-step concept of centralized group service would also be required for reasons previously mentioned. With both types of Centrex made available, the telephone company has full flexibility to plan in either direction depending on the economics of each case and the circumstances surrounding it.

The conversion of the standard conventional individual *701B* step-by-step PBX equipment to a centralized unit has been relatively simple and inexpensive. Fig. 11 shows the conventional *701* and Fig. 12 indicates the modification required to convert it to a centralized unit. The switching equipment is essentially the same with these exceptions: (1) the standard *701B* line-finder group is split to serve two 100-line groups; (2) the line-finder switch can be used in either the 100- or 200-line group; and (3) the use of flexible wiring has been increased to provide for varying service requirements. The line-finder unit was modified by supplying 20 subgroup relays instead of ten in order to serve two 100-line groups; the start-lead wiring was modified to permit the independent advance of two groups; two sets of commutator leads and alarms were provided and common leads were duplicated. By rearranging the equipment and by the substitution of wire spring-type relays for flat spring-type relays, these modifications are contained within the standard framework. The split is provided on an optional basis, so that by placing or cutting straps at the unit terminal strips, the grouping may be changed from two 100-line groups to one 200-line group or vice versa, respectively. The sleeve (*S* and *S1*) leads in the line-finder switch have been brought out on separate plug terminals. The wiring of the shelf unit determines whether the switch will hunt 200 or 100 lines. Complete freedom in the moving of switches between groups is therefore achieved.

Since, in a centralized PBX, each customer is cross-connected to the proper amount of equipment to meet his traffic needs, the switching and trunking equipment must be fully flexible. First and second selector banks are terminated at the distributing frame on a per-shelf basis. Line-finder, selector, and connector switches are also terminated at the distributing frame on an individual basis. Central office trunks, tie trunks, conference plans, traffic registers, etc., are also terminated there on an individual basis. Allocation to individual customers and grading of these equipments can be easily accomplished by cross connection, as can any subsequent changes.

The use of a common main frame, common power equipment, and in the case of "in dialing," common call-distribution equipment and common key-pulsing senders and links, not only reduce the investment and maintenance costs but eliminate the "getting started" cost that is required when PBX customers are served on an individual basis.

The *701B* Centrex system provides many of the service features of the no. 5 crossbar version, such as direct inward dialing, pushbutton console operation with key-pulsing completion, etc. However, since it is PBX equipment, it does not have the flexibility of the no. 5 concept, it cannot serve individual lines, and it must be trunked to and from a central-office unit. Outward identification through the centralized automatic message accounting operator is available rather than full automatic number identification, although it is hoped soon to have an economical arrangement for this feature. Nevertheless, the over-all long-range economies and improved service gained from the centralized *701* concept make it very attractive for those locations where no. 5 does not fit into the picture.

Fig. 13 shows the design effort required in developing the Centrex *701*. Most of the work was required to provide DID with transfer of calls to the proper customer's attendant.

The DID train is a 4-wire train, the fourth wire being used for visual signals to the attendant indicating to her the progress of the call. DID calls can be completed to PBX stations through the dial switching equipment or can be routed to a PBX attendant. If the directory listed number is dialed, the first digit received will advance the incoming first selector to the attendant level. The selector, by operation of a set of normal post springs, signals the incoming trunks to route the call to the PBX attendant. The attendant answers the call and completes it to the proper station through the dial switching equipment on a released-loop basis. Released-loop operation permits the attendant to reuse the incoming first selector to complete a call to any station, special position, or tie trunk to an associated PBX. When the called party answers, the attendant loop releases and is available for another call. In order to complete calls to stations and trunks, the attendant uses d-c key pulsing to a register sender which sends dial pulses to the switching equipment.

The call-distribution unit allots incoming calls, attendant trunk calls, and transfer calls to the associated PBX attendant position. A maximum of four PBX customers may be assigned to each call-distribution unit. The number of call-distribution units required will be determined by the number of customers and the number of incoming trunks and attendant's positions required by each customer.

Associated with each call-distribution unit are 24 trunk finders and 24 position finders. The trunks appear on the banks of the trunk finders, while the attendant loops appear on the banks of the position finders and are associated with one attendant's telephone set on a level. When a trunk requires connection to an attendant's position, the call-distribution unit starts a pair of trunk and position finders. The finders connect the trunk to an attendant loop circuit. The loop circuit gives an incoming call indication at the associated attendant's position. The trunk and position finders are divided into two groups, permitting one trunk and position finder in each group to operate simultaneously. The attendant connects to a register sender through a link circuit common to a maximum of 20 positions of one or more customers and 10 register senders. The register sender is arranged to send out dial pulses at the rate of 10 or 20 pulses per second depending upon the type of equipment into which it is pulsing.

A manual CAMP ON feature is provided for the attendant to "camp" on an incoming call on a busy line. This feature is a function of the connector and is activated by a positive potential applied to the sleeve of the switch train from the attendant's loop. When the customer is camped on a busy line, it remains on the line in a noncut-through position. As soon as the line becomes idle, the connector functions to ring the extension and then functions in a normal manner.

Two-way position trunks are key-terminated and provide a means for orig-

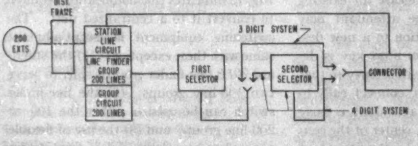


Fig. 11. Conventional 701B SXS PBX basic line group, 200 lines

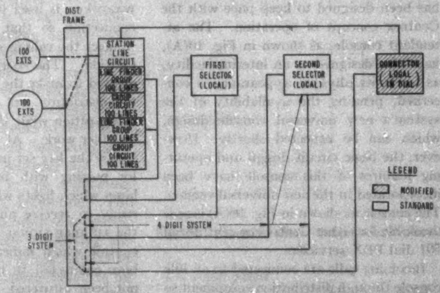


Fig. 12 (right). Centralized 701B SXS PBX basic line group, 100 lines

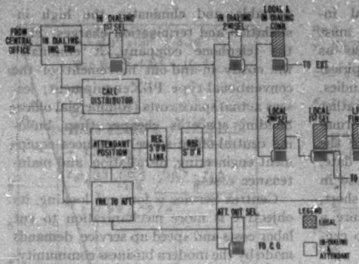


Fig. 13 (left). Centrex centralized 701B SXS PBX arranged for DID

inating calls in two directions. This feature is intended primarily to be used in connection with delayed long-distance calls and conference calls. It also provides for a six-button key telephone set to be used as a night station for receiving and completing calls after hours. When used with the night station, one selector associated with the two-way trunk is used by the night watchman or night attendant to complete the call to the requested station.

Practical Application of Centrex Service and Its Impact on the Bell System

The huge Grand Central City building "triggered" the development of no. 5 centralized customer group service and its companion, the *701B* step-by-step version. At the start of development, November 1959, the service date for the first installation (Grand Central City) was October 1962. The no. 5 crossbar equipment required for this location consists of a total of two complete marker entities (40,000 numbers) with four central-office designations. The extremely high calling rate in the Grand Central area makes it doubtful if more than 30,000 lines at the most can be served from these units, 15,000 in the Pan Am building itself and the other 15,000 in nearby buildings. Three nearby buildings, 277 Park Avenue, 280 Park Avenue, and 330 Madison Avenue, totaling well over 2,000,000 sq ft of rentable area, are already under construction for completion shortly after Pan Am. These will be served from the Pan Am unit. Some lower calling-rate individual lines or PBX trunks will probably be brought into the units in order to obtain the greatest efficiency from the equipment. The 0, 1, and 9 levels of each central-office designation will be available for this low calling-rate traffic, some of which will probably be residential individual lines.

Fig. 14. Irving Trust Company old operating room, replaced June 1960

This objective of October 1962 was no small order. Among other things, it called for negotiations with the owners covering the leasing of 35,000 sq ft of space on a long-term basis and for the provision of adequate floor strengths, ceiling heights, splicing chambers, cable entrances, beam clamps, and other items.

What looked like a very tight schedule in 1959, today appears easy because of the impact on the Bell System of transactions at Grand Central City. As a result, some 15 months prior to the original objective service date, the first no. 5 crossbar (centralized) Centrex office is already in service: the Wabash office in Chicago. Illinois Bell Telephone Company will have several others in service this year and by early 1962, including the O'Hare Airport installation. Bell of Pennsylvania has two in Philadelphia ready to cut and several others in the planning stage. The New York Telephone Company has at least five to be installed ahead of Pan Am: Bridge Street, Brooklyn; Idlewild Airport, Queens; RCA Communications Building, 60 Broad Street; Sperry Rand; and the New York State unit in Albany. There are more than 30 no. 5 crossbar centralized units in the Bell System scheduled for service by January 1, 1963. In the years ahead, each one of these Centrex units, and many more like them, will play a major role not only in giving their customers improved and speedier service, but in going a long way toward eliminating the costly in-and-out movement of the conventional type of PBX service.

The *701B* centralized concept will also prove its value many times over in the years ahead. Several installations of the *701B* Centrex noncentralized version have been made in the last year or so; most notable were the Time and Life Building, Union Carbide Corporation, and Metropolitan Life Insurance Company. The Irving Trust Company in-

Fig. 15. Irving Trust Company new operating room with pushbutton consoles

stallation in June 1960 was the first to get the new pushbutton consoles similar to those discussed.

A centralized *701B* installation is scheduled for early in 1962 at the Equitable Life Assurance Building at 393 Seventh Avenue. This will be another first, in that the existing 1,800-line conventional *701* unit now located in the customer's premises will be converted into a centralized Centrex unit and the space will be rented by the telephone company.

Before leaving the design discussion, it should be emphasized that this new type of telephone service must be viewed much differently, from an engineering standpoint, from anything previously done in the PBX equipment field. For instance, many of the concepts of central-office equipment engineering, including traffic, outside plant, and transmission, must now be applied to those centralized locations. The initial installation, as well as all additions, must be engineered on the same basis as central office units, based on empirical data (commercial forecasts, traffic studies, etc., for each centralized location) rather than specific tenant occupancy information. This is not only difficult for the sales people to obtain at an early date, but it is usually insufficient for use in proper centralized planning. In short, each centralized location will require a fundamental plan study similar to central office studies.

Conclusions

Centralized customer group service has many advantages for both the customer and the telephone company. For the customer, it gives speedier service; saves space which is becoming more valuable; and eliminates the high installation and termination charges. For the telephone company, it eliminates the costly in-and-out movement of the conventional-type PBX equipment; lessens actual space costs (commercial office-building space is cheaper than building central-office space); reduces equipment engineering, installation, and maintenance costs.

Centrex Service is forward looking, its objective is more mechanization to cut labor costs and speed up service, demands made by the modern business community. For example, the old Irving Trust Company multiple-cord 8-position switchboard, Fig. 14, has been replaced with the new Centrex units, Fig. 15. In view of this contrast one must conclude that there is a "new look" and a step forward made in the design of customer group telephone service for the modern business community.

Pole-Zero Techniques Applied to Voice-Frequency Telephone Lines

PAUL FLEMING, JR.
NONMEMBER AIEE

MOST TEXTBOOKS about transmission lines derive differential equations pertaining to physical phenomena and, after suitable solutions are found, immediately define or describe several parameters: (1) the propagation function and its components, the attenuation and phase shift functions; (2) the velocity of propagation; and (3) the characteristic impedance. Although others exist, the characteristics mentioned are of fundamental importance and interest, and hereinafter will be referred to in the collective sense as the basic parameters.

The propagation function for a transmission line with distributed constants is defined by the relation

$$\gamma=\sqrt{YZ} \tag{1}$$

$$=\sqrt{(G+j\omega C)(R+j\omega L)} \tag{2}$$

$$=\alpha+j\beta \tag{3}$$

The velocity of propagation is

$$V_p=\omega/\beta \tag{4}$$

and the characteristic impedance is

$$Z_0=\sqrt{Z/Y} \tag{5}$$

$$=\sqrt{\frac{(R+j\omega L)}{(G+j\omega C)}} \tag{6}$$

where R, L, G, and C are the magnitudes of the line parameters, expressed in appropriate units per length of line; α is the attenuation function; β is the phase shift function.

Routine substitutions and manipulations reduce equations 2 and 6 to

$$\gamma=\sqrt[4]{(RG)^2+\omega^2[(RC)^2+(LG)^2+\omega^2(LC)^2]}\;\underline{\Big/\frac{\tan^{-1}\frac{\omega[RC+LG]}{RG-\omega^2LC}}{2}} \tag{7}$$

$$Z_0=\sqrt[4]{\frac{(RG)^2+\omega^2[(RC)^2+(LG)^2+\omega^2(LC)^2]}{[G^2+\omega^2C^2]^2}}\;\underline{\Big/\frac{\tan^{-1}\frac{\omega[LG+RC]}{RG+\omega^2LC}}{2}} \tag{8}$$

Similar equations for α and β can be derived, each of which is equally complicated and unwieldy.

Both of the foregoing equations are useful, but it is difficult to develop from them an intuitive feeling about the behavior of transmission lines with changes in frequency or in the distributed constants. To evaluate either equation 7 or 8 requires a relatively large amount of straightforward but very tedious labor, with the result that few persons ever attempt such an investigation. Textbooks are of little assistance, because most of these in early pages make the simplifying assumption that the line is lossless, and is operated at a fairly constant frequency. Upon applying the first of these constraints, these relations hold:

$$\gamma=\sqrt{(j\omega C)(j\omega L)}=\sqrt{LC}\,\underline{/90^\circ} \tag{9}$$

$$Z_0=\sqrt{L/C}\,\underline{/0^\circ} \tag{10}$$

Unfortunately, telephone lines are not lossless, nor is the voice-frequency range of 200 to 4,000 cycles merely a narrow band of 3,800 cycles; instead, it is 4.3 octaves, or 1.3 decades of change. The basic parameters ordinarily undergo the greatest changes in this range; consequently, the analysis of telephone lines is quite rudimentary unless calculations are made at several frequencies. The discussions which follow will demonstrate methods for determining quickly the basic parameters as functions either of frequency or of the distributed constants. Further, interesting conclusions can be

---

Paper 61-113, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 7, 1960; made available for printing February 27, 1961.

PAUL FLEMING, JR., is with Automatic Electric Laboratories, Inc., Northlake, Ill.

This paper is based on "The Calculation of Transmission Effectiveness of Subscribers' Lines in Small Telephone Exchanges," a Master's Thesis submitted to the University of Arkansas, Fayetteville, Ark., in 1953.

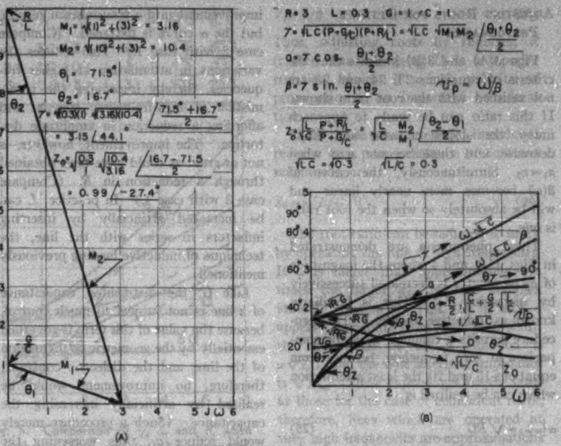


Fig. 1. (A) Plot of $a_1 = -1$, $a_2 = -10$, on complex frequency plane, and (B) variations in line parameters for conditions in (A)

drawn about other phenomena, such as the correlation between the impedance at zero and at infinite frequencies.

Complex Plane Plots

COMPLEX FREQUENCY[2]

A more powerful study can be made if $j\omega$ is replaced in equations 1 through 6 with the term P, where P represents a complex frequency of form

$$P = a + j\omega \qquad (11)$$

After the substitution mentioned, equation 2 is transformed into

$$\gamma = \sqrt{(G+pC)(R+pL)} \qquad (12)$$

$$\gamma = \sqrt{LC(p+G/C)(p+R/L)} \qquad (13)$$

$$= \sqrt{LC(p-a_1)(p-a_2)} \qquad (14)$$

and 6 becomes

$$Z_0 = \sqrt{\frac{R+pL}{G+pC}} \qquad (15)$$

$$= \sqrt{\frac{L}{C}\frac{(p+R/L)}{(p+G/C)}} \qquad (16)$$

$$= \sqrt{\frac{L}{C}\frac{p-a_2}{p-a_1}} \qquad (17)$$

ROOTS OF TRANSMISSION-LINE EQUATIONS

The respective roots of equations 13 and 16 are numerically equal, and exist at the points $p_1 = a_1 = -G/C$ and $p_2 = a_2 = -R/L$. In the equation for γ, equation 13, both a_1 and a_2 are zeros, whereas a_1 is a pole of the characteristic impedance expression. Because of the dual roles of a_i in the subsequent analysis, these terms will be referred to hereinafter merely as the roots.

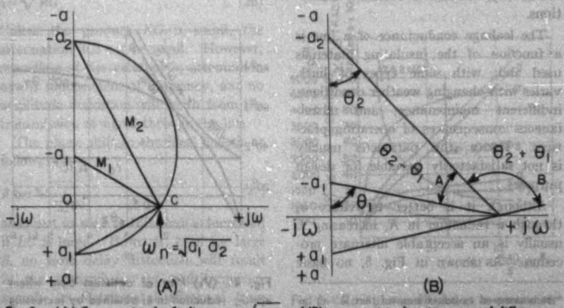


Fig. 2. (A) Construction for finding $\omega_n = \sqrt{a_1a_2}$, and (B) construction showing sum and difference of angles θ_1 and θ_2

COMPLEX-FREQUENCY PLANE

As shown in Fig. 1(A), the roots of equations 13 and 16 can be plotted on a complex-frequency plane. For any frequency, the magnitude and phase of the propagation function is

$$\gamma = \sqrt{LC}\sqrt{M_1M_2}\angle\frac{\theta_1+\theta_2}{2} \qquad (18)$$

where the M_i are the magnitudes of complex variable factors. Similarly, the characteristic impedance is

$$Z_0 = \sqrt{L/C}\sqrt{M_2/M_1}\angle\frac{\theta_2-\theta_1}{2} \qquad (19)$$

Also of interest are the relations

$$\alpha = |\gamma|\cos\frac{\theta_1+\theta_2}{2} \qquad (20)$$

$$\beta = |\gamma|\sin\frac{\theta_1+\theta_2}{2} \qquad (21)$$

Specific solutions for the propagation functions and for the characteristic impedance are shown in the illustrations. The magnitudes of M_1, M_2, θ_1, and θ_2 can be determined either by analytical techniques or by measuring directly from the complex plane. Unless extreme precision (better than 1% or 2%) is required, measurement is much faster and yields creditable results. Both axes of the complex plane must be linearly calibrated; otherwise, improper relations between M_1, M_2, θ_1, and θ_2 will be obtained. In particular, semilog or double-log paper is unsuitable for the reasons just given, and also because no origin exists.

CONSTRUCTIONAL AIDS

As will be described in a later section, the frequency

$$\omega_n = \sqrt{a_1a_2}$$

is important because certain symmetries occur about this point. By simple proportion, the distance oc, Fig. 2(A), can be shown to be equal to ω_n. For any frequency, the angle A of Fig. 2(B) is $\theta_2 - \theta_1$, and B is $\theta_2 + \theta_1$.

THE ATTENUATION FUNCTION, OPTIMUM CRITERIA

The attenuation function, α, describes the diminution of the signal as it progresses along the line from input to

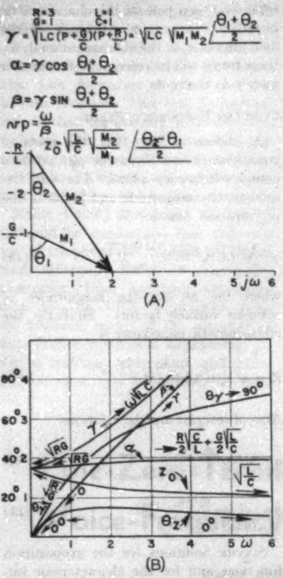


Fig. 3. (A) Plot of $a_1=-1$, $a_2=3$, on complex-frequency plane; reduction in a_2 obtained by increasing L; and (B) variations in line parameters for conditions in (A)

output terminals. If α is independent of ω, signals of any frequency within the pass band of interest will be transmitted with equal effectiveness. This desirable condition is stated by

$$d\alpha/d\omega=0 \tag{22}$$

for frequencies between approximately 200 and 4,000 cps (cycles per second).

The term β represents the phase shift of the signal—with respect to the phase of the applied voltage—as a function of its propagation along the line. It is preferable that β be zero or else proportional to frequency, the first condition being merely a special case of the latter. More explicitly

$$\beta=K\omega+0 \tag{23}$$

(It is not sufficient that $d\beta/d\omega=K$. It is necessary that the linear $B(\omega)$ function pass through the origin.)

The telephone line should present a constant impedance, both to its load and its source, to minimize the effects of mismatching. This requirement implies the following:

$$dZ_0/d\omega=0 \tag{24}$$

ADJUSTING ROOTS FOR OPTIMUM PERFORMANCE

Figs. 3(A) and 3(B) indicate that the criteria of equations 23, 24, and 25 are not satisfied with the root ratio shown. If this ratio can be made to approach unity, the slopes of $\alpha(\omega)$ and $Z_0(\omega)$ decrease, and these become zero when $a_1=a_2$. Simultaneously, the curve of $\beta(\omega)$ becomes more nearly linear and will be absolutely so when the root ratio is one.

These phenomena are demonstrated in Figs. 1, 3, and 4 where the magnitude of a_2 parameter is decreased successively by increasing L.* (This is familiarly known as inductive loading.) This procedure permits a common basis for comparison at zero frequency, because from equations 13 and 20, the least attenuation which can be realized is

$$\alpha|_{\omega=0}=\sqrt{RG} \tag{25}$$

Hence, if R and G remain unchanged, the transmission effectiveness at zero frequency also remains unchanged, regardless of variation in L or C.

Also, a_2 can be reduced by decreasing R. Fig. 5 shows the γ, α, and β curves of Fig. 1 (case 1, $a_2=-10$), of Fig. 3 (case 2, $a_2=-3$), and of a third transmission line (case 3), in which $a_2=-3$ (the same value used in case 2), is obtained through a reduction in R. As is readily apparent, case 3 represents a major improvement over case 2 in transmission effectiveness, because, at all frequencies, the $\gamma(\omega)$ curve is lower and flatter, and the $\alpha(\omega)$ curve is more linear than the corresponding curves in case 2.

Thus, it is desirable that, if possible, the root ratio be reduced by decreasing R, because the inherent (d-c) transmission of the line is improved thereby. Normally, however, R cannot be reduced indefinitely because of economic limitations.

The leakage conductance of a line is a function of the insulating materials used and, with some types of lines, varies with changing weather conditions, indifferent maintenance, and miscellaneous consequences of operating practices. Hence, this parameter usually is not satisfactorily variable for design purposes.

Although it is better to reduce a_2 through a reduction in R, increasing L usually is an acceptable alternate procedure. As shown in Fig. 5, no basic improvement in d-c transmission results, but the α curve is flattened. (Compare case 2 with case 1.) This reduces the variations in attenuation at higher frequencies, thereby improving the transmission of high-frequency signals, and affording a reduction in amplitude distortion. The improvement, however, is not as great as could have been obtained through a reduction in R. (Compare case 3 with case 1.) In practice, L can be increased artificially by inserting inductors in series with the line, the technique of inductive loading previously mentioned.

Like G, the distributed capacitance of a line is not subject to much control, because the value of this term is governed essentially by the geometric configuration of the line, and the dielectric constant; therefore, no improvement would be realized by artificially increasing the capacitance. (Such a procedure merely would reduce a_1, thus worsening the situation.)

Although it would be advantageous, no effective method is available for decreasing the magnitude of this parameter. Once it is decided to use 24-

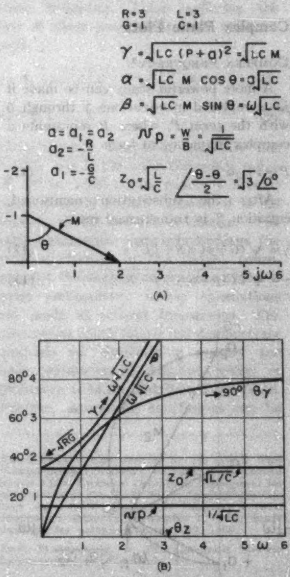


Fig. 4. (A) Plot of optimum case where $a_1=a_2$; reduction in a_2 obtained by increasing L; and (B) variations in line parameters for conditions in (A)

---

* It is assumed in this discussion that $a_2>a_1$, because this is the usual case. Should $a_1>a_2$, the same conclusions are drawn if G is substituted for R, and C for L.

CASE	R	L	G	C	$\frac{R}{L}$	$\frac{G}{C}$	$\sqrt{RG}$	$\sqrt{LC}$
1	3.0	0.3	1	1	10	1	$\sqrt{3}$	$\sqrt{0.3}$
2	3.0	1.0	1	1	3	1	$\sqrt{3}$	$\sqrt{1}$
3	0.9	0.3	1	1	3	1	$\sqrt{0.9}$	$\sqrt{0.3}$

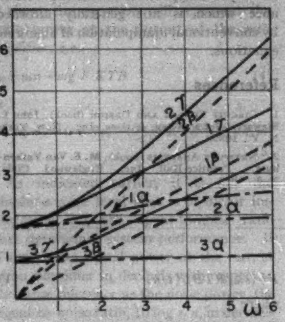


Fig. 5. Comparison of γ, α, and β, when a_1 is changed from −10 to −3 by increasing L (case 2), and by reducing R (case 3), the latter showing superior characteristics

LC cable, for example, the exchange engineer must accept the capacitance which is characteristic of that cable.

THE OPTIMUM CASE (DISTORTIONLESS LINE)

The two roots of the propagation function (and characteristic impedance) equation should coincide if absolute optimum conditions are required. As shown in Fig. 4, such an ideal situation results in a constant α, constant Z_0, constant velocity, and a phase shift proportional to frequency.

In the ideal case, the attenuation for any frequency is reduced to that of the d-c case, or

$$\alpha=\sqrt{RG} \qquad (26)$$

Unless the product *RG* is small, the attenuation will not be small. However, regardless of the value of α, attenuation now is independent of frequency, and no amplitude distortion will result from the transmission of signal through the line.

The phase shift in the ideal case is as follows:

$$\beta=\omega\sqrt{LC} \qquad (27)$$

and, just as with α, this term is increased if *LC* is large. However, even for large β, no time delay distortion will result in the transmission of multifrequency signals. A detrimental effect of large *LC* is the resulting low velocity, because as noted previously in equation 4,

$$V_p=\omega/\beta \qquad (28)$$

Thus, adjusting roots by reducing *R* (rather than increasing *L*) is a superior technical choice, if a high velocity of propagation is an important consideration.

Finally, the characteristic impedance reduces to a pure resistance of constant magnitude, or

$$Z_0=\sqrt{L/C}\,\underline{/0^\circ} \qquad (29)$$

The root ratio can be reduced to unity by operating the line at high frequencies. If ω is larger than approximately ten times the larger a_1, the two complex variable vectors of Figs. 1, 3, and 4 appear to emanate from a common point, the magnitude of the γ function is $\omega\sqrt{LC}$, and the characteristic impedance is $\sqrt{LC}$. These expressions are the same as those for the case of coincident roots; therefore, lines which are operated at very high frequencies are approximations of the ideal distortionless line which was discussed in the preceding paragraphs.

Applications and Interpretations

The principles just described were employed to produce the curves of Fig. 6 and 7. In both cases, the frequency

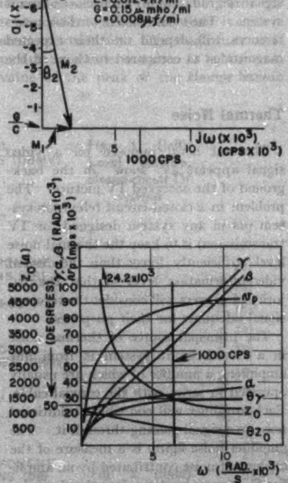


Fig. 6. Root plot and variation in parameters for 109–135 steel-conductor open-wire facility

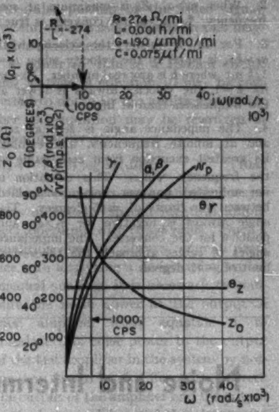


Fig. 7. Root plot and variation in parameters for 24-gage paper-insulated cable

$\omega_n=\sqrt{a_1a_2}$ is quite low, being below or near the lower limit of the voice-frequency band. This is typical of all commonly used telephone-wire facilities.

For most telephone lines, a_1 is about 20; therefore, the angle θ_1 approaches 90 degrees whenever $\omega>(10)(20)=200$, or approximately 30 cps. The angle θ_2 is nearly zero for frequencies less than 0.1 a_2; hence, for open-wire facilities ($a_2=8$ to 10×10^3), the second root begins to have significant effects at about 100 to 150 cps. For this type of facility, the second root contributes an angle of approximately +45 degrees near the middle of the voice band; therefore, the impedance angle usually is between −20 degrees and −30 degrees at 1,000 cps.

In contrast, cables manifest second roots of from 85,000 (19 gage) to 220,000 (26 gage). Again, when $\omega<0.1\ a_2$ (or about $14\text{–}15\times10^3$ cps for 19 gage, 3,500 cps for 26), the second root contributes little to changes in basic parameters. Thus, for cables operated in the voice range (particularly the smaller gages), the impedance angles are very nearly −45 degrees. For similar reasons, the angle of the propagation function also is about 45 degrees in this range. Hence, for most cables, one finds that $\alpha\approx\beta$.

Numerous other conclusions can be formulated from the methods described, but space permits mention of only a few of them:

1. When plotted upon double-log paper, α, β, V_p, and Z_0, all exhibit double-mirror symmetry about the frequency $\sqrt{a_1a_2}$.

2. When $a_2>a_1$, Z_0 is maximum at zero frequency; for $a_1>a_2$, the converse is true.

3. For any frequency, the velocity when $a_2=ka_1$ is equal to the velocity when $a_2=1/k\,a_1$; where k is any real number. Except for $k=1$, the velocity always is lower at low frequencies than at high.

4. The impedance angle is zero at zero and at infinite frequencies, and exhibits its greatest excursion from zero value at the frequency $\sqrt{a_1a_2}$. This deviation is not unlimited, however, being constrained between the limits of ±45 degrees. The angle always is negative for $a_2>a_1$, and positive for the converse. The impedance angles of cables approach the theoretical limit of −45 degrees.

Conclusions

The techniques presented herein introduce no new parameters, but instead provide a simple way of evaluating terms frequently found in the literature. Detailed knowledge of the frequency behavior of a transmission line sometimes is more interesting than useful, except in a secondary sense, where such information can be employed very effectively in designing filters or equalizers which are to be inserted in voice-frequency lines; in calculating reflection losses at junctions of dissimilar lines; and for other applications wherein several octaves of frequency change are involved. Perhaps most important of all these considerations, interpretation of pole-zero plots develops an appreciation of line performance which is not generally afforded by conventional manipulation of algebraic equations.

References

1. Circuit Theory and Design (book), John L. Stewart. John Wiley & Sons, Inc., New York, N. Y., 1956.

2. Network Analysis (book), M. E. Van Valkenburg. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1955.

Noise and Intermodulation Problems in Multichannel Closed-Circuit Television Systems

C. A. COLLINS
NONMEMBER AIEE

A. D. WILLIAMS
NONMEMBER AIEE

IN THE past decade there has been a rapid growth in communication facilities designed to transmit television (TV) signals relatively short distances over coaxial cables. These systems handle six or more video signals which amplitude-modulate (AM) r-f (radio-frequency) carriers in the 8- to 220-mc (megacycle) range. The associated audio signals frequency-modulate (FM) r-f sound carriers whose rest frequencies are 4.5 mc away from their associated picture carriers. These systems are used for industrial, educational, and entertainment purposes, and descriptions of many of them have appeared in the literature.[1-3] This paper will discuss two problems that affect the design of any closed-circuit TV system: thermal noise and intermodulation products created by amplifier nonlinearity.

These factors limit the ultimate useful length of a closed-circuit TV system. Optimum design will occur when noise and nonlinearity objectives are just met at the extreme ends of the system.

The design objectives for the limitation of these interferences have been based on the results of subjective tests. In the broadcast field the Bell System is furnishing combinations of intercity TV transmission facilities which, taken together on a transcontinental basis, are designed to meet these objectives. In the field of closed-circuit TV it is usually possible to allocate the entire allowable cross-country signal degradation to each closed-circuit system. The visual effects of these interferences will depend on their expected magnitudes as compared to those of the desired signals.

Thermal Noise

Thermal noise imposed on a video signal appears as "snow" in the background of the observed TV picture. The problem in a closed-circuit television system (as in any system designed for TV transmission) is to keep the thermal-noise level sufficiently lower than the desired video-information level so that the majority of viewers will consider the received signal adequate for their purposes.

The principal source of thermal noise in a closed-circuit system is from line amplifiers, a number of which are usually operated in series with the coaxial cable. Each amplifier will contribute additional noise to a signal passing through it. The amplifier-noise figure is a measure of the amount of noise contributed by an amplifier.[4] It may be defined as the degradation in signal-to-noise ratio incurred by a signal as a result of its transmission through the amplifier, i.e.,

$$F=(s_{\text{in}}/KTB)/(s/n) \qquad (1)$$

where

F = noise figure
s_{in} = input-signal power
KTB = available input thermal-noise power
s = available signal power at amplifier output
n = available thermal-noise power at amplifier output

For the case of an amplifier with matched input and output impedances the gain of the amplifier is $g=s/s_{\text{in}}$, so that equation 1 may be rewritten as:

$$n=gFKTB \qquad (2)$$

or, expressed in decibels (db) above unit power,

$$10 \log n=10 \log g+10 \log F+10 \log KTB \qquad (3)$$

In words, the output-noise power, 10 log n, of an amplifier is the sum of its gain in db (G), its noise figure in db (F_{db}), and the available input-noise power, i.e.,

$$10 \log n=G+F_{\text{db}}+10 \log KTB$$

It should be noted that the thermal noise at the output of an amplifier actually includes input cable thermal noise of strength $g\ KTB$ and amplifier thermal noise of strength $(F-1)\ g\ KTB$. In these systems F is in the 8–10 range and the amplifier-noise contribution is the controlling factor in total system noise. For this reason, system noise is considered as if it were only a function of the amplifiers in the system.

A closed-circuit system does not consist of a single amplifier, but rather a series of amplifiers separated by lengths of coaxial cable. There is a build-up of noise power

---

Paper 61-181, recommended by the AIEE Broadcasting Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted November 30, 1960; made available for printing March 16, 1961.

C. A. Collins and A. D. Williams are with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

at amplifier outputs. With all amplifiers equally spaced and operating with the same gain, each one will contribute an equal amount of noise to the system. Thus, if equation 2 represents the noise contribution of a single amplifier, m times the equation will represent the noise power at the output of the mth amplifier, i.e.,

$$n_m = mn = mg\ F\ KTB \tag{4}$$

or

$$10\log n_m = G + F_{db} + 10\log KTB + 10\log m \tag{5}$$

As indicated earlier, it is not the absolute noise level that is of major importance, but the signal-to-noise ratio that determines system performance. If S represents signal power at the output of an amplifier in decibels with respect to the same reference as the noise power, the signal-to-noise ratio, $10\log s/n$, in decibels at the output of the mth amplifier is:

$$10\log(s/n)_m = 10\log s - 10\log n_m$$

$$10\log(s/n)_m = S - G - F_{db} - 10\log KTB - 10\log m \tag{6}$$

However, since $G = S - S_{in}$, where S_{in} = amplifier input-signal power in decibels with respect to the same reference as the noise power, equation 6 may be written:

$$10\log(s/n)_m = S_{in} - F_{db} - 10\log KTB - 10\log m \tag{7}$$

The noise margin in a system may be defined as the number of decibels by which computed signal-to-noise performance, $10\log(s/n)_m$, exceeds a signal-to-noise objective $(10\log s/n)_0$.

$$\text{Noise margin} = 10\log(s/n)_m - (10\log s/n)_0$$

$$\text{Noise margin} = S_{in} - F_{db} - 10\log KTB - 10\log m - (10\log s/n)_0 \tag{8}$$

The r-f signal-to-noise ratio objective $(10\log s/n)_0$ to be used in this equation is derived from a base-band signal to weighted noise ratio requirement of 54 db, where the base-band signal to weighted noise ratio is expressed as $20\log_{10}$ (peak-to-peak video-signal magnitude/weighted rms noise magnitude). The translation of the 54-db video requirement into an r-f requirement of 48 db assumes a flat r-f noise spectrum and an AM r-f carrier received on a conventional TV receiver.

In a typical system $[10\log s/n]_0 = 48$ db, $F_{db} = 10$, $10\log KTB = -107.57$ dbm (db referred to 1 milliwatt) in a 4-mc bandwidth. Equation 8 may be written as:

$$\text{Noise margin} = S_{in} - 10 + 107.75 - 10\log m - 48$$

$$\text{Noise margin} = S_{in} + 49.75 - 10\log m \text{ (db)} \tag{9}$$

At the end of an ideal system the noise margin may be zero and equation 9 may be rewritten as:

$$S_{in} = -49.75 + 10\log m \text{ (dbm)} \tag{10}$$

The minimum input necessary to meet system-noise objectives for various numbers (m) of amplifiers in tandem can be determined using this equation.

In closed-circuit systems carriers are normally measured across 75 ohms in terms of dbmv, where zero dbmv is equivalent to 1 millivolt. The magnitude of a modulated carrier is taken to be the rms value of the peak of the carrier wave. This peak occurs normally during the sync-pulse portion of the video signal. System objectives are based on this magnitude specification.

A power in dbm in a 75-ohm circuit such as that in equation 10 may be converted to dbmv by the addition of 48.75; then the equation may be rewritten in terms of dbmv as:

$$S_{in} = -1 + 10\log m \text{ (dbmv)} \tag{10A}$$

These equations are applicable to systems of equally spaced identical amplifiers, operating with equal gains. Such a condition does not always occur in practice. In analyzing a design using varied unequal spacings it is necessary to compute the noise power at the output of each amplifier using equation 3, to translate this noise power to the output of the last amplifier in the system by noting any signal-level differences between the output of the amplifier and the output of the last amplifier, and then to sum up on a power basis the noise contribution of each amplifier to the total noise power. This total noise power is then compared to signal power at the output of the last amplifier to determine the signal-to-noise ratio of the system.

The use of a level diagram such as the one shown in Fig. 1 is desirable since computations based on its signal magnitudes will reflect the effects of misequalization, level-control action, and temperature variation on system performance.

This computation of noise performance can be made for various temperatures of operation and at various frequencies though usually system-noise performance will be worst at the highest temperature and highest frequency of operation. In determining amplifier spacing, input signals must be kept high enough so that the computed signal-to-noise performance of the system will meet or exceed the design objectives.

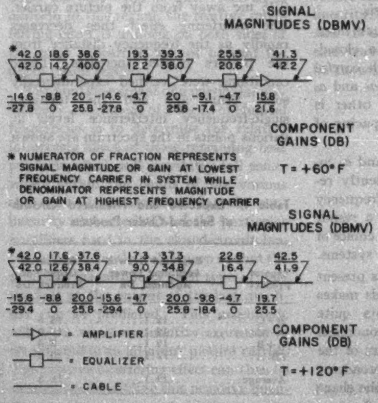

SIGNAL MAGNITUDES (DBMV)
*42.0/42.0 18.6/14.2 38.6/40.0 19.3/12.2 39.3/38.0 25.5/20.6 41.3/42.2
-14.6/-27.8 -8.8/0 20/25.8 -14.6/-27.8 -4.7/0 20/25.8 -9.1/-17.4 -4.7/0 15.8/21.6
COMPONENT GAINS (DB)
* NUMERATOR OF FRACTION REPRESENTS SIGNAL MAGNITUDE OR GAIN AT LOWEST FREQUENCY CARRIER IN SYSTEM WHILE DENOMINATOR REPRESENTS MAGNITUDE OR GAIN AT HIGHEST FREQUENCY CARRIER
T = +60° F
SIGNAL MAGNITUDES (DBMV)
*42.0/42.0 17.6/12.6 37.6/38.4 17.3/9.0 37.3/34.8 22.8/16.4 42.5/41.9
-15.6/-29.4 -8.8/0 20.0/25.8 -15.6/-29.4 -4.7/0 20.0/25.8 -9.8/-18.4 -4.7/0 19.7/25.5
COMPONENT GAINS (DB)
T = +120° F
= AMPLIFIER
= EQUALIZER
= CABLE

Fig. 1 (left). Level diagram, closed-circuit TV systems

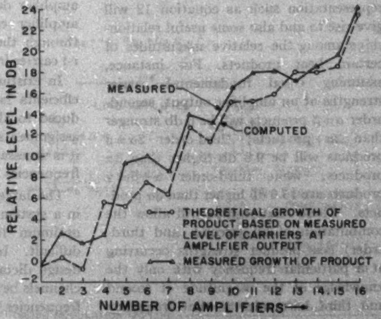


Fig. 2 (right). Growth of A+B−C product for 16 amplifiers

Amplifier Nonlinearity Problems

Just as thermal-noise considerations will impose a lower limit on the allowable signal-input magnitude to an amplifier, amplifier nonlinearities will impose an upper limit on the output-signal magnitude. The basic video transmission requirements that must be invoked in considering amplifier nonlinearities are those on: (1) single-frequency interference, and (2) cross-modulation (XM) effects.

In a video-transmission path the presence of an extraneous single frequency of a relatively low magnitude can produce a noticeable and distinctive pattern on a TV picture. Such single frequencies are generated in nonlinear amplifiers as beat products between the r-f carriers of the various channels on the system. Cross modulation is another form of interference arising from amplifier nonlinearity. It is one in which the modulation on one carrier is imposed on another. An analysis of the first type of interference will lead to an understanding of the second.

The manner in which these unwanted frequencies are generated can be seen by considering the instantaneous input-output characteristic of an amplifier as:

$$e_{out}=a_1e_n+a_2e_{in}+\ldots a_ne_{in}^n+\ldots \qquad (11)$$

Practically, over the range of signal magnitudes in which amplifiers operate in these systems, it is sufficient to consider only the first three terms in this expression. Thus:

$$e_{out}=a_1e_{in}+a_2e_{in}^2+a_3e_{in}^3 \qquad (12)$$

The expansion of equation 12 for even such a simple input signal as:

$$e_{in}=A\cos\omega_at+B\cos\omega_bt+C\cos\omega_ct \qquad (13)$$

yields over 30 output products in addition to the desired input signals. These additional products occur as second- and third-order beat products. The expansion indicates the type of products that a representation such as equation 12 will give rise to and also some useful relationships among the relative magnitudes of certain beat products. For instance, assuming equal fundamental carrier strengths at an amplifier output, second-order $\alpha\pm\beta$ products will be 6 db stronger than 2α products; third-order $2\alpha\pm\beta$ products will be 9.6 db higher than 3α products; while third-order $\alpha+\beta\pm\gamma$ products are 15.6 db higher than 3α products. These relationships will allow the computation of all (second- and third-order) beat-product strengths occurring at a particular frequency with only the knowledge of the magnitude of second and third harmonics at that frequency with respect to the fundamental magnitudes.

In an actual system there may be as many as six video carriers, six audio carriers, and two pilot tones fed into an amplifier. A method of computing the location of all products resulting in such a situation is discussed later.

A fundamental system consideration, in addition to the output magnitudes of these frequencies from a single amplifier, is the rate of growth of the these products in a string of amplifiers in tandem. It has been found in other multichannel coaxial-cable systems that certain third-order products will tend to add on a voltage basis (20 log *m*), whereas all other third-order products and all second-order products will, on the average, add on a power basis (10 log *m*).[5] The expected rate of growth of products in a closed-circuit system has been confirmed by measurements on a system operated by the Chesapeake and Potomac Telephone Company of Maryland in Hagerstown. The results of these measurements can be seen in Fig. 2 and Table I. Fig. 2 shows the computed levels for a third-order $A+B-C$ interference product based on the assumption of voltage addition of that product, (20log *m*). The fairly close agreement of these results indicates the general correctness of the assumption of voltage addition for this product. Table I contains the results of measurements and computations made on second-order products which were assumed to add on a power basis. It can be seen from the table that based on the overall average of these measured products, power addition of these products is indicated. These results confirm the theory. Third-order products of the $A+B-C$ and $2A-B$ type will tend to add on a voltage basis, whereas second-order products will tend to add on a power basis.

There are two ways in which these interferences can be minimized in a closed-circuit system. One is through careful amplifier design to reduce the a_2 and a_3 amplifier coefficients and the other is through the use of judicious spacing of r-f carrier frequencies.

In existing amplifiers the a_2 and a_3 coefficients have not been sufficiently reduced to allow indiscriminate frequency assignments for a system. As a result, it is necessary to make a careful choice of frequencies in closed-circuit TV systems.

The large number of products present in a system of up to six channels makes optimum frequency assignments quite difficult. In the first place, economical design dictates that at least part of the channels be carried at standard broadcast frequencies. The remainder of the channels may be carried in the sub-very-high-frequency range from 25 to 50 mc with provision for conversion to standard frequencies at the receiving points. Whatever channel assignments are used, it is desirable that the beat products formed by the interaction of their various carriers fall outside the chosen channel bands or within the bands at points where they will cause the least interference with received pictures.

It is obvious that the computation of the frequencies of the numerous beat products resulting from the combinations of the video, sound, and color carriers for a number of channel combinations would be a formidable task if undertaken by hand. Consequently, use was made of an International Business Machines Corporation *704* computer. A program was written that would calculate all the products and arrange them in numerical order together with a code denoting the sources of these products for given channel assignments. This list could then be examined for interferences near or within each channel band.

Where the beat product interferences fall outside the channel band there is no problem. There will always be many products, however, that will appear within the band, and the amount of interference that can be tolerated varies quite widely depending upon the part of the channel spectrum in which it appears. In a channel carrying only black and white picture information the most sensitive portion of the spectrum is at, or near, the frequency of the picture carrier (called 0 frequency in Fig. 3). There is considerably less sensitivity at a frequency about 15 kc away from the picture carrier but the interference becomes quite large again at a frequency about 0.75 mc away from the picture carrier. The interfering effect then decreases rapidly as the frequency distance from the picture carrier increases. These conditions were verified experimentally some years ago and the tolerable single-frequency interference levels at various points in the spectrum are shown

Table I. Expected and Measured Growth of Second-Order Products

Type of Product	Expected Growth Assuming Power Addition, Db	Measured Growth of Product, Db
B−C	12.1	15.6
A−C	12.1	15.9
A+B	12.1	3.5
2B	12.1	10.2
Average	12.1	11.3

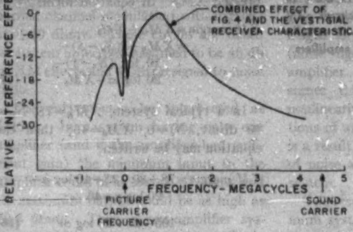


Fig. 3 (left). Relative effect of single-frequency interferences, black and white only

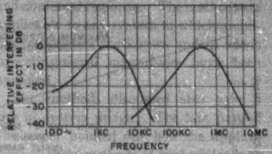


Fig. 4 (right). Tolerable single-frequency interference levels

approximately in Fig. 4. These levels would be uniform on both sides of the picture carrier. However, when the interference is viewed on a standard TV receiver the vestigial side-band nature of the receiver response further reduces the effect of all frequencies appearing on the vestigial side of the picture carrier. The combined effect of the tolerable interference levels shown in Fig. 4 and the vestigial nature of the receiver response have been plotted together in Fig. 3 which shows the relative effect of single interferences occurring at various frequencies in the spectrum of a channel carrying only black and white.

In a channel carrying color the region of the color subcarrier, approximately 3.6 mc above the picture carrier, becomes an additional sensitive point and the tolerable interference levels at various points near it also can be taken from Fig. 4. In this case, however, subjective tests have indicated that the relative effect of an interference occurring directly at the color subcarrier is about 4 db less than one appearing directly at the picture carrier. Since the pass band of color receivers is uniform in this region, the relative interfering effect is uniform on both sides of the color subcarrier, except beyond the upper cutoff region.

The final assessment of a particular preliminary assignment of a series of channel frequencies is to determine the relative interfering effect of those remaining beat products which will inevitably fall somewhere within the bands of the various channels. It is assumed that the nonlinearity characteristics of the particular amplifiers used in the closed-circuit system are sufficiently well known so that the probable magnitude of each type of beat product at the output of the last amplifier can be predicted. This magnitude will be a negative quantity expressed in db referred to an adjacent picture carrier. The relative interfering effect can then be determined by adding this negative quantity in db to a second minus quantity chosen from Fig. 3, depending upon the location of the interference product in the frequency spectrum of the particular channel. The system objective is that the sum of these quantities should be at least −60 db. This standard is related to the interference which subjective tests have indicated could be detected in a receiver connected to the output of the last amplifier in the system.

The minimizing of the interfering effect of carrier beat products by judicious frequency selection leaves third-order XM interference as a controlling factor in system performance. Third-order XM is a unique effect because it is an interference that cannot be moved about with respect to the carrier being interfered with, even though the interfering signal be moved throughout the pass band of the amplifier. It is an effect in which the modulation on one carrier is imposed on a second carrier. It is caused by a third-order $A+B-C$ type of product. This can be seen by an examination of Fig. 5 which shows how the modulation on one carrier appears to be transferred to a second carrier as a result of amplifier nonlinearity. The illustration is based upon AM of a carrier by a single-frequency sine wave. The modulated wave can be shown to consist of the sum of three sine waves, one occurring at the frequency of the unmodulated carrier ω_b and the other two $\omega_b \pm \delta$ separated from it by an amount equivalent to the frequency of the modulating wave δ. At the input to the amplifier the signal shown in Fig. 5 consists of two modulated carriers, ω_b and ω_a. At the output of the amplifier, and because of the nonlinearity in the amplifier, there are additional frequencies present. Among these frequencies are third-order beat products between the modulated carrier ω_b, its side bands $\omega_b \pm \delta$, and ω_a. Two of these beats are: $\omega_a + (\omega_b \pm \delta) - \omega_b = \omega_a \pm \delta$. These beats appear as side bands to the ω_a carrier thus giving an apparent additional modulation to this carrier. Notice that these beat products occur as side bands on the ω_a carrier and are independent of the location of the ω_b carrier.

This figure shows the XM effect in terms of carriers amplitude modulated by single-frequency sine waves. In actual system operation the carriers are modulated by relatively complex video signals. The XM effect in this case can also be explained by the mechanism of Fig. 5. Instead of the simple spectrum of a carrier plus two sine-wave side bands, the video-modulated carrier could be shown by Fourier analysis to contain many sine-wave side-band components. Each of these components would react in the same way as the side bands, in the simple case, to produce a corresponding component on other carriers going through

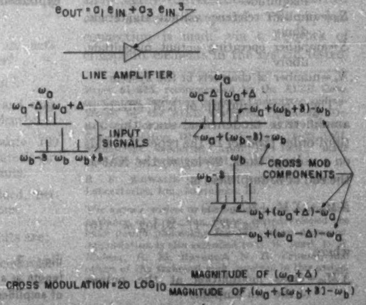

$e_{OUT} = a_1 e_{IN} + a_3 e_{IN}^3$
LINE AMPLIFIER
ω_a
$\omega_a - \Delta$
$\omega_a + \Delta$
INPUT SIGNALS
$\omega_b - \delta$
ω_b
$\omega_b + \delta$
$\omega_a + (\omega_b + \delta) - \omega_b$
$\omega_a + (\omega_b - \delta) - \omega_b$
CROSS MOD COMPONENTS
$\omega_b + (\omega_a + \Delta) - \omega_a$
$\omega_b + (\omega_a - \Delta) - \omega_a$
CROSS MODULATION = 20 $\log_{10} \frac{\text{MAGNITUDE OF } (\omega_a + \Delta)}{\text{MAGNITUDE OF } (\omega_a + [\omega_b + \delta] - \omega_b)}$

Fig. 5. Cross modulation as an A+B−C type of product

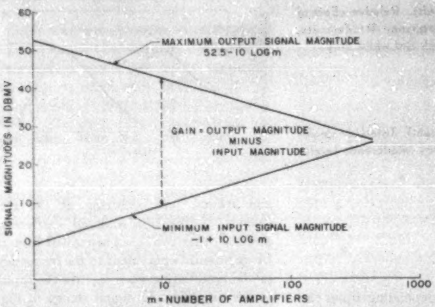


Fig. 6. Operating magnitudes in closed-circuit TV system with equally spaced amplifiers

the amplifier. The result would be the transfer of complex video modulation among carriers. The simpler single-frequency case is easier to visualize and measurements are often made using sine-wave modulation though the results and their interpretation are independent of the modulating signal.

XM may be measured as the ratio of the magnitude of desired signal modulation on a carrier to that of the transferred modulation on a carrier. Expressed in decibels, XM may be defined as:

$$XM = 20 \log_{10} \frac{\% \text{ normal modulation}}{\% \text{ imposed modulation}} \qquad (14)$$

In a single amplifier the amount of cross modulation is a function of the output magnitude of the carriers and the number of modulated carriers in use in the system. Thus, at the output of a single amplifier and assuming that the interfering effects of many channels will combine on a power basis:

$$XM = XM_R + 2(S_R - S) - 10 \log (N_c - 1) \qquad (15)$$

where

XM_R = cross modulation at the output of a single amplifier measured for two carriers at the S_R reference output magnitude
S_R = amplifier reference output magnitude, dbmv
S = amplifier operating output magnitude, dbmv
N_c = number of channels in the system

There is a build-up of XM effects in amplifiers in tandem and since this is a third-order product of the type that adds on a voltage basis, 20 log m, the XM at the end of m amplifiers is:

$$XM_m = XM_R + 2(S_R - S) - 10 \log (N_c - 1) - 20 \log m \qquad (16)$$

where

XM_m = cross modulation at the output of the mth amplifier
m = number of amplifiers in tandem

and the other quantities are as defined in equation 15.

Subjective tests on allowable flat crosstalk between video circuits have indicated that interferences of this type should be 58 or more db below desired signal magnitudes.[6] Since r-f cross modulation is similar to base-band video crosstalk, 58 db has been taken as a system XM objective.

This objective obtains for systems operating without a common synchronizing source on all channels. Under these conditions the lack of synchronism between independent video signals can cause a relative motion between interference and signal that is most disturbing and leads to the 58-db requirement. A means of reducing the effect of XM between channels is to operate the system with all video signals from a common synchronizing source. Under such a condition a more liberal XM objective might be applied. Such a limit has not been subjectively determined, but a tentative value of about 46 db has seemed in order.

The XM margin of the system may be defined as the number of decibels by which the computed cross modulation, expressed in db, below the wanted signal exceeds the XM objective, similarly expressed XM_0. In equation form:

$$\begin{aligned} XM \text{ margin} &= XM_m - XM_0 \\ &= XM_R + 2(S_R - S) - 10 \log (N_c - 1) - 20 \log m - XM_0 \qquad (17) \end{aligned}$$

In a typical system, $XM_R = 78$, $S_R = 40$ dbmv, $N_c = 6$, $XM_0 = 46$; then the equation may be written:

$$\begin{aligned} XM \text{ margin} &= 78 + 80 - 2S - 10 \log 5 - 20 \log n - 46 \\ &= 105 - 2S - 20 \log m \qquad (18) \end{aligned}$$

With the XM margin reduced to zero at the end of the system, equation 18 may be rewritten as:

$$S = 52.5 - 10 \log m \text{ (dbmv)} \qquad (19)$$

and, just as equation 10 gave minimum input-signal magnitudes necessary to meet thermal-noise objectives, this equation gives maximum output signal magnitudes necessary to meet linearity objectives for various numbers of amplifiers in tandem. The linearity characteristics of closed-circuit TV amplifiers are such that the use of signal magnitudes determined by this type of equation will result in amplifier operation within its normal overload limits.

Equations 14–19 apply to a system with equally spaced, identically operated amplifiers and can only provide a guide to the design of an actual system. Cross-modulation effects in a proposed system can best be determined by using actual expected output magnitudes as found in a diagram such as Fig. 1. Such a determination will take into account misequalization, level-control action, and temperature variations.

The expressions for minimum input-signal magnitudes, equation 10, as determined by thermal-noise considerations, and maximum output-signal magnitudes as set by XM effects, equation 19, can be combined to give a measure of system performance by a given amplifier. These

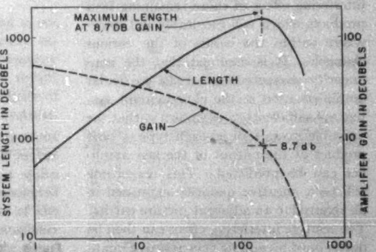


Fig. 7. System maximum length as a function of number of amplifiers and amplifier gain

equations are shown plotted in Fig. 6 for an amplifier having a 10-db noise figure and cross modulation 78 db below normal channel modulation with carriers at +40 dbmv. These curves are based on system objectives assumed to be 46 db on XM effects and 48 db on signal-to-noise ratio.

The curves may be interpreted as follows. In a system having only one amplifier (and if the amplifier had sufficient gain) the minimum input to the amplifier could be as low as −1 dbmv and its maximum output could be as high as 52.5 dbmv. Thus, a one-amplifier system could be as long in length as 53.5 db. For a two-amplifier system the minimum input would be +2 dbmv and the maximum output signal would be 49.5 dbmv. Each amplifier could use a gain of 47.5 db and maximum system length would be 95 db (versus 53.5 db for a one-amplifier system).

This procedure might be continued indefinitely (or at least until amplifier input- and output-signal magnitudes were nearly equal) to give a measure of maximum system length as a function of number of amplifiers in the system. It would seem, as indicated in the examples, that the lower the gain required of the amplifier the greater the maximum possible system length. The result of this computation of maximum length and amplifier gain is shown in Fig. 7. System maximum length does not increase indefinitely but rather is a maximum when amplifier gain reaches 8.7 db. This maximum length is a function of the properties (noise figure and nonlinearity) of the amplifier. The 8.7-db-gain figure assumes that amplifier-noise figure and nonlinearity performance are not functions of amplifier gain. This gain figure is a result of the different rates of growth of noise and nonlinearity effects in the system.

The result of this analysis is that maximum system length from a given amplifier is obtained when the amplifier is used with 8.7 db of gain. A more general result is that a system's noise and nonlinearity performance can be improved if amplifier spacing is reduced. The useful limit to this reduction is with amplifiers which are spaced 8.7 db apart. Reduced amplifier spacing implies increased system first cost as well as operating expenses. The relative worth of improved performance versus increased costs involves matters outside the scope of this paper.

Conclusions

This paper has described methods for evaluating the performance of r-f closed-circuit TV systems with respect to interferences caused by amplifier thermal noise and nonlinear response. The methods are applicable to any system in which amplifier parameters are known and in which operating-signal magnitudes have been specified.

The methods discussed also have a use in system design. The last part of the paper pointed out how the application of these methods of analysis leads to the conclusion that system noise and nonlinearity performance can be improved by reducing amplifier spacing. As noted, the limit to this reduction in spacing occurs with amplifiers spaced 8.7 db apart.

References

1. Closed-Circuit Network for Educational Television, W. C. Warman. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 76, Sept. 1957, pp. 440–44.

2. A Transmission Design for Closed Circuit Educational Television, W. L. Wall. *Ibid.*, vol. 78, Sept. 1959, pp. 340–46.

3. Redstone Arsenal Closed-Circuit Educational Television System, J. P. Argo, I. N. Howell. *Ibid.*, vol. 79, May 1960, pp. 185–90.

4. Noise Figures of Receivers, H. T. Friis. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 32, July 1944, p. 419.

5. The L3 System-Design, C. H. Elmendorf, R. D. Ehrbar, R. H. Klie, A. J. Grossman. *Bell System Technical Journal*, New York, N. Y., vol. 32, July 1953, p. 792.

6. Observer Reaction to Video Crosstalk, A. D. Fowler. *Journal*, Society of Motion Picture and Television Engineers, New York, N. Y., vol. 57, Nov. 1951, pp. 416–24.

Transmission Network of an Electronic Crosspoint PABX

R. F. KOWALIK
NONMEMBER AIEE

THE PURPOSE of this paper is to discuss the transmission network of a private automatic branch exchange (PABX) which is undergoing operational testing at Automatic Electric Laboratories, Inc., in Northlake, Ill. It was designed to use space-division transmission and time-division common control.

The system had its origin in a 100-line private automatic exchange (PAX)[1] built at the laboratories for internal use only. The PABX, in so far as the transmission facilities are concerned, was to be an expanded copy of the PAX, except for generally improved circuits and the addition of circuitry necessary to handle supervisory signals, conference access, and trunks.[2] Its transmission facilities include the switching network and various line circuits.

Outstanding features of the switching network are:

1. Selected crosspoint elements are activated by a common system-marker.
2. Silicon p-n-p-n diodes are used.
3. Split junctors permit ringing and ringback tones to be applied independently.
4. The network will accommodate 100 subscriber lines, 20 trunks, 4 operator's circuits, and 6 conference-access circuits.
5. Thirty-six junctors are provided, permitting 36 simultaneous connections.

Primary features of the line circuits are:

1. Busy tone and dial tone are controlled at the line circuit.
2. Outpulsing on trunks is accomplished by using tone signaling between line circuits.
3. Strapping options on line circuits are provided to restrict subscribers from trunk access and to designate which subscribers are to receive night calls.

Crosspoint Device

In a space-division system, each conversation must be transmitted by a separate physical connection, unlike time- or frequency-division schemes, which allow many conversations to be carried over relatively few paths. The physical connection is made via a network of crosspoint elements in the system tested.

Paper 61-825, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted April 10, 1961; made available for printing April 25, 1961.

R. F. Kowalik is with Automatic Electric Laboratories, Inc., Northlake, Ill.

The author wishes to acknowledge the advice and guidance of J. G. Van Bosse, who developed the basic design philosophy for the system discussed. Appreciation is also extended to C. H. Insel, D. B. Moberg, R. M. Ravanesi, N. H. Zeumer, and others of the technical staff involved in this assignment.

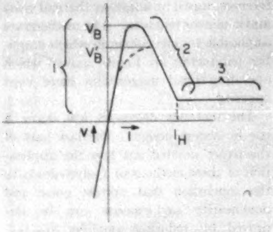


Fig. 1. Voltage–current characteristics of a silicon p-n–p-n diode

The obvious question is: What is a crosspoint and what does it do? In answer, consider the proverbial black box having a number of input and output terminals. Crosspoint elements would be those devices which could be connected, one between every input and every output in such fashion that, by activating the correct element, any input could be connected to any output. Crosspoint elements are basically single-pole single-throw switches. Their ideal characteristics are that the OFF state exhibits an infinite impedance between terminals whereas the ON state exhibits a zero impedance.

Devices which could be used as crosspoint elements include relay contacts, gas tubes, and a large assortment of semiconductors. In this system, the device used is the silicon p-n–p-n diode, a 4-layer diode, whose voltage–current characteristics are shown in Fig. 1.

Three separate regions have been designated on the curve. Region *1* is OFF, and when operated in this region, the device exhibits an impedance greater than 1 megohm. Region *2* is a transition region which exhibits a negative-resistance characteristic. Region *3* is ON. When operated in this region, the device shows an impedance of a few ohms, in series, with a constant direct-voltage drop of approximately 1 volt.

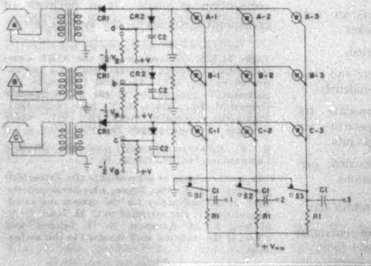
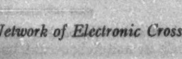


Fig. 3 (left). A 3 by 3 crosspoint matrix

Fig. 4 (below). Voltage waveform at cathode of p-n–p-n diode



Explanation of these voltage-current characteristics involves visualizing the p-n–p-n device as an n-p-n and a p-n-p transistor, connected back-to-back, the outside layers functioning as emitters; see Fig. 2.

The ON state is reached when the sum of the alphas of both transistors exceeds unity. The alphas depend both on current density and on the reverse potential across the center junction in both transistors (avalanche multiplication).

Starting in the OFF state, the sum of the low-current alphas is well below unity. Increasing the voltage across the device, the alphas increase because of avalanche multiplication until their sum reaches unity at the breakdown voltage V_B. Starting at the ON state, the high-current alphas are sufficiently large to insure their sum being greater than unity. Decreasing the current through the device, however, decreases both current-sensitive alphas until their sum just equals unity at the holding-current level, I_H. If the current is reduced further, the device turns off because the sum of the alphas drops below unity.

The function of the switching network is to shift the device from region *1* to region *3*, and maintain the device at some operating point in region *3* throughout a call, after which it must return the device to region *1*.

To turn the device on it is subjected, temporarily, to a voltage greater than the breakdown voltage. The actual breakdown voltage of the device is somewhat dependent on the rise time of the applied voltage, the so-called "rate effect." The dotted line in Fig. 1 illustrates how the characteristics change during application of a fast pulse. To turn the device off, the network diverts the supplied holding current from the crosspoint.

Crosspoint Network

Having specified a crosspoint device, it becomes necessary to arrange a group of them in such an array that they can be controlled independently.

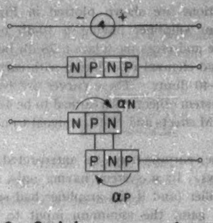


Fig. 2. Symbol, structure, and transistor analogy of p-n–p-n diode

Fig. 3 shows a typical matrix of crosspoints. Each horizontal input lead terminates in a transformer winding, representing a subscriber's line transformer. Vertical output leads each terminate in a resistor connected to a positive voltage source. The switch on the crosspoint side of the resistor permits this point to be switched to ground. The capacitor *C1* provides an output to some other circuit. The other switch shown appears on each subscriber's link and serves to supply a negative potential.

The circuit operates in this way: Assuming subscriber *B* is to be connected to output *2*, the first action is the marking of the subscriber link. Switch *b* is operated, causing link *B* to go negative as shown in Fig. 4. The slow rise time is effected by capacitor *C-2* and is provided to minimize false firing resulting from the "rate effect."

Three p-n–p-n diodes are marked with $-1/2\ V_B$ on their cathodes. Since the connection is to be to line *2*, switch *2* is operated, permitting the potential on line 2 to rise toward V_{BB}. As soon as p-n–p-n diode *B-2* sees its full breakdown, it turns ON. The capacitor below the link charges towards the positive potential, determined by the voltage divider.

As soon as the potential on link *B* rises above ground, diode *CR1* starts conducting and link *B* is held close to ground, as determined by the direct-voltage drop of the transformer winding and *CR1*. Switch *b* is then returned to its original position, causing capacitor *C2* to charge positively, reverse-biasing *CR2* and electrically removing the link-marking circuit from the link. Assuming resistor *R1* has been selected small enough to provide the holding current for the p-n–p-n diode, it will remain in its ON state and the connection is established.

With a 1-volt drop in the transformer winding and *CR1*, the potential on link *B*

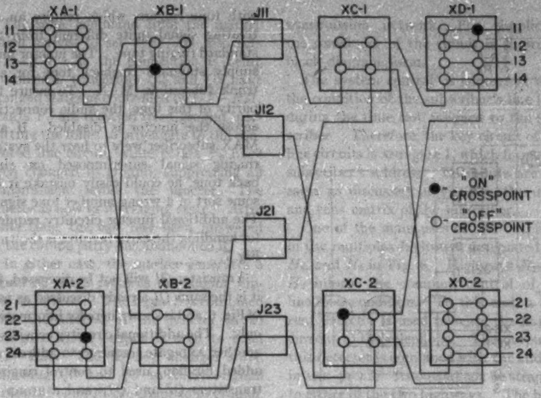


Fig. 5. An 8-subscriber 4-stage crosspoint network

will be a positive 1 volt. Adding the drop across the p-n-p-n diode, the potential on line *2* will be 2 volts positive. The independence of each crosspoint from every other is proved when a second call is set up in the same matrix without affecting the existing call.

Consider now that the subscriber at *A* is to be connected to line *3*. Line *A* is marked negatively to $-1/2\ V_B$ as before. The crosspoint *A-2* will see the voltage difference between $-1/2\ V_B$ and the direct potential on line *2*, plus the peak-positive excursion of the voice-frequency signals on this line. (The sum of these two positive voltages is small compared with $+1/2\ V_B$; thus crosspoint *A-2* is in no danger of firing spuriously.) The connection cycle continues with switch *3* operating, and the new connection is completed. An examination of all the transient and steady-state potentials reveals that at no time was there any danger of a crosspoint firing falsely.

At the termination of a call, a crosspoint is returned to its OFF state by operation of the switch which supplied the positive firing potential. Returning this switch to its ground signal diverts the holding current of the crosspoint to ground, and it reverts to its OFF state.

The next step in forming a network is to connect a multitude of these matrices together. As in the original PAX, a 4-stage network, symmetrical about a center junctor, is used here. The junctor serves to join the two halves of the network, and supplies the positive half of the crosspoint breakdown voltage. Fig. 5 illustrates a 4-stage network for an 8-subscriber system. Note that each subscriber appears on both sides of the network. The system control requires that all calls originate on the left side of the network, *XA* matrix, and terminate on the right side, *XD* matrix. The two inner matrices are designated *XB* and *XC*, left and right of center, respectively.

The fact that each subscriber's line equipment has two connections into the transmission network permits the subscriber to be connected as both a calling and a called party simultaneously. This has many benefits; it permits such features as chain conference calls, trunk transfer, operator intercept, etc. The fully implemented PABX would have this complement of crosspoint matrices: ten *XA* and *XD*, and six *XB* and *XC*. Thirty-six junctors join the two halves of the network. Each *XB* matrix has a junctor common to each *XC* matrix.

A subscriber address (number) is a 2-digit number and is determined by which *XA* and *XD* matrix the subscriber is in, and his position in the matrix. Thus a subscriber located in *XA*, matrix 2, in the third level would have his address formed by tens-digit 2 and units-digit 3, which is simply 23. Similarly, the junctor address is a 2-digit number based on the number of the *XB* and *XC* matrices it connects; the first digit refers to the *XB* matrix, the second to the *XC*. Thus, a junctor joining *XB* matrix 1 to *XC* matrix would be addressed as 12.

As can be seen from the illustration, any subscriber can be connected to any other subscriber, using any junctor. The task of selecting which junctor is to be used for a particular call is performed by the marker[3] in an operation called route selection.

Fig. 6 illustrates a portion of the actual circuits used in this system, and their operation can be seen from the illustration. Assume the marker receives a request to connect subscriber 23 to subscriber 11, and that the route search has found the call may use *J-12*.

Initially, the junctor *J-12* is idle while flip-flop *F-F—J-12* is in state *0* and supplies base current to *Q1*, which is in saturation. Transistor *Q1* carries currents from $+V_{BB}$ through resistors *R1* and diodes *CR1*, so the potential at P_A is close to ground. Transistors *Q2* are off and the potentials at P_B and P_C of the idle links are also close to ground.

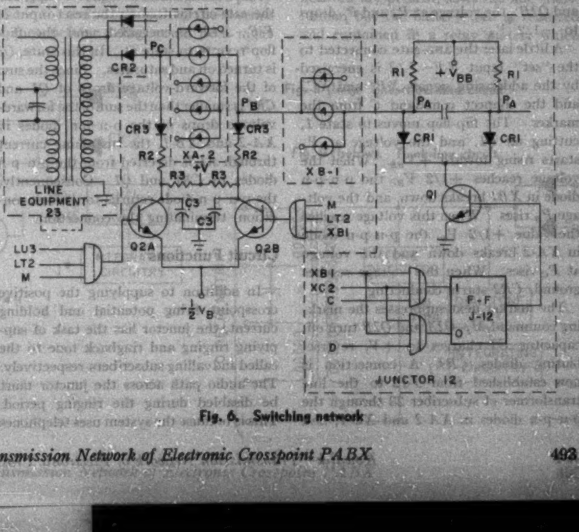


Fig. 6. Switching network

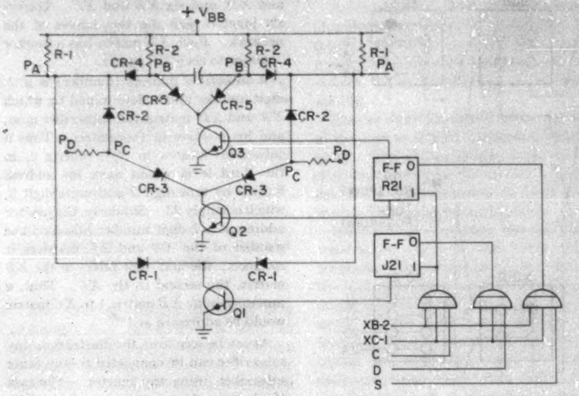


Fig. 7. Junctor circuit

The potential $+V$ maintains a reverse bias on diodes *CR3*.

The first step in initiating a connection involves marking the selected links of the route. Every link marker contains a transistor *Q2*; capacitor *C3*; diode *CR3*, and resistors *R2* and *R3*. Transistors *Q2* are driven by AND-gate circuits, which in turn are fed from marker outputs. Since subscriber 23 is to be connected to the left-hand side of the network, marker outputs *LT2* (left tens-2) and *LU3* (left units-3) are energized. Since the connection is to be routed over *XB1*, the marker output *XB1* is energized. During the marking time, the marker-output *M* is also energized. This activates the AND-gates and saturates *Q2A* and *Q2B*; the voltage at P_B and P_C drops to $-1/2\ V_B$.

A little later the AND-gate connected to the "set" input *F-F—J-12* is energized by the addressing signals *XB1* and *XC2* and the connect command *C* from the marker. The flip-flop moves to state *1*, cutting off *Q1*, and the voltage at P_A starts rising toward $+V_{BB}$. When the voltage reaches $+1/2\ V_B$, the p-n-p-n diode in *XB1* breaks down, and the voltage P_B rises. When this voltage reaches the value $+1/2\ V_B$, the p-n-p-n diode in *XA-2* breaks down and the voltage at P_C rises. When this voltage reaches ground, *CR2* starts conducting.

The marker next suppresses the marking command *M; Q2A* and *Q2B* turn off, capacitor *C3* charges to $+V$, reverse-biasing diodes *CR3*. A connection is now established from *J-12* to the line transformer of subscriber 23 through the p-n-p-n diodes in *XA-2* and *XB-1*, and the link markers have been electrically removed from the links. At the same time, a similar process initiates the connection in the right-hand half of the network between *J-12* and the called subscriber. The 4-layer diodes in both branches receive sustaining current through resistor *R1*. Voice-frequency currents are carried from one side of the network to the other through capacitor *C* in the junctor circuit.

This situation continues until the marker receives a DISCONNECT order, which is accomplished by the address of the junctor that held the connection. In this case, marker outputs *XB1* and *XC2*, addressing junctor *J-12*, are energized. When the disconnect signal *D* appears, the AND-circuit feeding the reset input of *F-F—J-12* is energized, and the flip-flop reverts to state O. In this state, *Q1* is turned on and saturates. Since the sum of the forward-voltage drops of *Q1* and *CR1* is smaller than the sum of the forward-voltage drops of the p-n-p-n diodes in *XA-2* and *XB-1*, the sustaining current through *R1* is diverted from the p-n-p-n diodes to *CR1* and *Q1*. Consequently, the p-n-p-n diode returns to the OFF condition, terminating the connection.

Circuit Functions

In addition to supplying the positive crosspoint firing potential and holding current, the junctor has the task of supplying ringing and ringback tone to the called and calling subscribers, respectively. The audio path across the junctor must be disabled during the ringing period. This is because the system uses telephones with tone ringers, which require an activating signal quite different from the standard ringing tone. The junctor must supply standard ringback tone for any trunk subscriber, MAX. To insure the purity of this tone, the audio connection across the junctor is disabled. If any MAX subscriber were to hear the system ringing signal superimposed on ringback tone, he could easily mistake it for some sort of a wrong-number tone signal. The additional junctor circuitry required for handling the extra features is shown in Fig. 7.

Transistor *Q1* will not be discussed, for it is the same *Q1* already discussed as part of Fig. 6, and is shown here for information only. The additional circuitry consists of another AND-gate, required to address an added flip-flop, used to control ringing; transistors *Q2* and *Q3*; and a group of diodes and resistors. Ringing tone is applied at point P_D on the side of the junctor connected to the called side of the network. Ringback tone is applied to the same point on the other side of the junctor. Ringing tone has a peak-to-peak maximum amplitude which is less than 2 volts, while ringback tone is much smaller. The tones have in common the fact that they are superimposed on a +10 direct-voltage level.

Placing the tones on the line is accomplished by diverting them from ground to the junctor output. With transistor *Q2* saturated, diodes *CR-3* are forward-biased, and the tones appearing at points P_C are clamped close to ground. To make the tones available at P_A, transistor *Q2* is cut off, removing the ground from diodes *CR-3* and permitting points P_C to rise towards +10 volts. This, then, forward-biases diodes *CR-2*, and the tones are available to the crosspoint network.

Preventing the tones from being transferred across the junctor is accomplished by reverse-biasing diodes *CR-4*. Normally, when it is desirable to disable the audio connection across the junctor, points P_A will have a positive potential in excess of +2.5 volts. By saturating transistor *Q3*, the potential at points P_B are clamped close to ground, and diodes *CR-4* are reverse-biased. A requirement here is that the value of resistors *R-2* be sufficiently small that the diodes *CR-4* will operate in a linear portion of their characteristics when they are required to carry audio signals.

The over-all operation of the tone gates is detailed next. Initially, in the idle state, transistor *Q2* is saturated, transistor *Q3* is cut off, and flip-flop *R21* is in its reset state.

The same command from the marker which arrives to set *F-F—J21* also sets *F-F—R21*. As the connection is being established, transistors *Q2* and *Q3* have changed states as controlled by *F-F—R21*. Points P_C are released from ground, permitting the tones to propagate to the connected line circuit, while points P_B have been clamped to ground, suppressing all transmission across the junctor. The junctor is now in its ringing state. After some time, the call has been answered or the calling party has abandoned it.

In either case, the marker generates a signal *S*, which requires the ringing flip-flop to be reset. The signal *S* is combined with the address of the junctor in an AND-gate whose output is used to reset *F-F—R21*. The flip-flop changes state, returning *Q2* and *Q3* to their original states. Points P_C are clamped to ground, reverse-biasing diodes *CR-2*. Points P_B are released from ground and rise until diodes *CR-4* become forward-biased. Points P_B will settle out at some voltage related to the potential at points P_A plus the drop across diodes *CR-4*. The tones are now not coupled to the talking path, and the subscribers are free to speak.

Fig. 8 illustrates a typical subscriber-line circuit. The line circuit in this system has several functions in addition to the normal one of providing the termination for the subscriber's line loop. The circuit gives loop supervisory information to the subscriber logic, furnishes dial and busy tone as required, and provides strapping options required to place designated subscribers in special categories.

The resistors, capacitor, and transformer form a standard battery feed system for the subscriber's telephone. The two diodes, *CR-1* and *CR-2*, provide output connections to the two sides of the transmission network. For simplicity, the remainder of the circuit is shown in block-diagram form.

The system control[9] is concerned with the condition of the subscriber's line loop during the time slot assigned to the subscriber. Therefore, the key circuit of the line circuits is AND-gate 1, which forms the subscriber's address. Its inputs are the same as discussed previously—the units and tens matrix of the subscriber.

One of the main uses of this output is in the multiplex highways designated H_1, H_2, and H_3 in Fig. 8. Highways H_1 and H_2 inform the common control of the line-loop condition. To this end, a current detector is used to sense the flow of current to the telephone. The output of this detector is combined with the address in AND-gate 2. Its output can be strapped to either of the two highways. The highway not receiving this output is fed the output of AND-gate 1. This strapping option is used as will be explained.

If a subscriber is permitted to dial a trunk call, he is strapped to modulate his H_1 highway; if he is restricted from this feature, he is strapped to H_2. The common control can thus detect that a subscriber is dialing, whether or not he is restricted, and, by appropriate timing logic, the digit dialed. Furthermore, if no output appears on either highway, the common control can conclude that no line equipment occurs at this time slot. This permits the number of subscriber lines in use to be changed simply by adding or removing subscribers' line circuits from the system. The common control automatically senses if a particular address is equipped with a line circuit.

Highway H_3 is used to designate a party who is to receive calls during the absence of the operator. This could be a night watchman, night foreman, etc. For this feature, the output of AND-gate 1 is strapped to the H_3 output terminal.

Dial tone is applied by means of a tone gate in the line circuit. The common control, during the time slot of a subscriber, generates signal *D* which actuates an integrating circuit connected as the input to the subscriber's tone gate. The integrator sustains the gate in the ON condition until the next time this subscriber's time slot appears; the signal is then either repeated or absent. The repeated signal maintains the tone on the line until such time as the common control fails to supply *D*. When busy tone is required, signal *D* is interrupted at the proper rate, and the subscriber accepts the interrupted dial tone as busy tone.

To dial out over a trunk, the system requires that the dial pulses of the telephone be converted to bursts of a tone signal, which are transmitted to the line circuit of the trunk where they are reconverted to d-c loop pulses. When a nonrestricted subscriber dials the code number assigned to signify that a trunk is desired, the control network provides the calling line circuit with signal *H*. This signal is gated with the output of the current detector in AND-gate 3. The succeeding dial pulses become actuating signals for the tone-signaling gate, and tone bursts are sent out as required.

The trunk-line circuit, Fig. 9, is quite similar to a subscriber-line circuit, for, as far as the transmission network is concerned, the trunk is just another subscriber. As in the standard-line circuit, the two diode outputs to the transmission network are present, as is the line transformer, but the battery feed is removed and contained in a relay matrix which

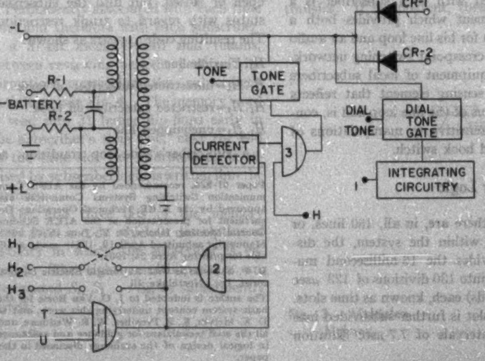
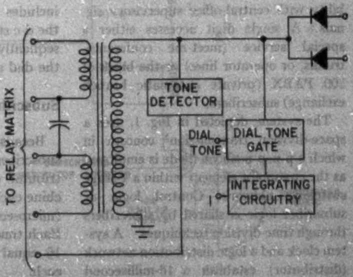


Fig. 8 (left). Subscriber-line circuit

Fig. 9 (below). Trunk-line circuit



connects the electronic system to the outside trunk lines. The addressing AND-gate is not present, for no time-slot information is developed at this circuit. The dial- and busy-tone signals control is the same as in the standard circuit. Busy tone is required for an incoming trunk call. That portion of the highway information necessary for the trunks is developed in the relay matrix.

The unique circuit on the trunk-line equipment is the tone detector, which consists of an amplifier feeding a tuned tank circuit. Output of the tuned circuit is rectified and used to trigger a Schmitt trigger circuit, whose output is coupled into a relay-driver circuit in the relay matrix. The relay pulses out on the trunk lines, and is in a sense an extension of the subscriber's dial.

Transmission Tests

To date, extensive measurements have not been made on the over-all transmission quality of the system. The 3-decibel bandpass has been measured to extend from 230 to 50,000 cycles per second. Calculations predict a crosstalk level of approximately −50 decibels. Measurements have not been made, but there is no audible crosstalk.

References

1. 100-Line Electronic Private Automatic Telephone Switchboard, John G. Van Bosse. *Automatic Electric Technical Journal*, Northlake, Ill., vol. 7, no. 2, Apr. 1960, pp. 50–57.

2. Features of an Electronic Crosspoint PABX, J. G. Van Bosse, J. DeCicco. *AIEE Transactions* (see pages 471–74 of this issue).

3. Logical Control of an Electronic Crosspoint PABX, R. P. Sanders. *Ibid.* (see pages 496–501.

Logical Control of an Electronic Crosspoint PABX

R. P. SANDERS
ASSOCIATE MEMBER AIEE

RECENTLY COMPLETED, after 15 months' work by members of Automatic Electric Laboratories, Inc., is a new electronic telephone switching system which will accommodate 100 local subscribers and 20 trunks emanating from an electromechanical exchange.[1] Ten of the trunks are PBX (private branch exchange) trunks and ten are 2-way dial trunks. Four operator lines have been established, not only as intermediate terminations for the PBX trunks, but also to provide special services for local subscribers. A 6-position meet-me conference facility is reinforced by a progressive or chain conference feature. Provisions have been made for trunk transfer, trunk restriction, night service, and operator intercept on unequipped lines.[2]

The exchange is a 2-wire facility that makes use of the standard telephone dial. Standard dial tone, ringback tone, and busy tone are used to insure compatibility with central-office supervisory signals. A single digit accesses either a special service (meet-me conference, trunks, or operator lines) or the block of 100 PABX (private automatic branch exchange) subscribers.

The system, depicted in Fig. 1, uses a space-division transmission[3] concept in which a p-n-p-n silicon diode is employed as the crosspoint element within a 4-stage switching network. Control logic or subscriber logic is shared by subscribers through time-division techniques. A system clock and a logic distribution network (distributor) establish a 16-millisecond machine cycle and subdivisions thereof to regulate the multiplexing operation.

Pertinent information concerning each subscriber is retained within a ferrite-core memory during the period when subscriber logic is not at the subscriber's disposal. The marker is a logic network that is stationary with respect to the time reference established by the distributor; it is therefore capable of reviewing the status of each subscriber on a one-at-a-time basis. The marker conveys information from one subscriber to another and performs the function from which it derives its name: that of marking the path through the crosspoint switching network that is to be used for a pending connection.

Associated with each subscriber is a line equipment which provides both a termination for his line loop and an audio link to the crosspoint switching network. The line equipment of local subscribers includes a sensing element that reflects the d-c status of the line loop and is, consequently, sensitive to manipulations of the dial and hook switch.

Subscriber Logic

Because there are, in all, 130 lines, or subscribers, within the system, the distributor divides the 16-millisecond machine cycle into 130 divisions of 123 μsec (microseconds) each, known as time slots. Each time slot is further subdivided into 16 equal intervals of 7.7-μsec duration each.

In forming time slots, the distributor generates 13 units-pulses (denoted U_{D1}, U_{D2}, U_{D3} . . . U_{D9}, U_{D10}, U_{DX}, U_{DY}, U_{DZ}) of 123-μsec length and recurring sequentially every 1,600 μsec, and 10 tens-pulses (designated T_{D1}, T_{D2}, . . . T_{DO}) of 1,600-μsec duration, which recur every 16 milliseconds or, in other words, every machine cycle. A complete units-pulse train is contained within each tens-pulse. The coincidence of a units-pulse and a tens-pulse designates a time slot. Time slot 68, for instance, would be generated by the simultaneous occurrence of T_{D6} and U_{D8}.

Each subscriber is permanently assigned a time slot in accordance with Fig. 2, whereupon his status is analyzed by the subscriber logic and reviewed by the marker. By interrogating the distributor, the marker readily discerns subscriber identity.

Three separate multiplex highways (H_1, H_2, and H_3) inform the subscriber logic of conditions existing within the subscriber's line circuit. H_1 and H_2 are generated in such a way as to reflect not only the line-loop condition, whether open or closed, but also the subscriber status with regard to trunk restriction. The resulting code appears as shown:

$H_1 . H_2$ = idle line
$\bar{H}_1 . H_2$ = unrestricted subscriber off hook
$H_1 . \bar{H}_2$ = restricted subscriber off hook
$\bar{H}_1 . \bar{H}_2$ = unequipped line

A subscriber's line-loop condition, as

Paper 61-826, recommended by the AIEE Communication Switching Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted April 10, 1961; made available for printing April 24, 1961.

R. P. Sanders is with Automatic Electric Laboratories, Inc., Northlake, Ill.

The author is indebted to J. G. Van Bosse for the basic-system concept underlying this work, and to D. K. Melvin, K. E. Prescher, W. R. Wedmore, and all the staff technicians for guidance and assistance in logical design of the exchange discussed in this paper.

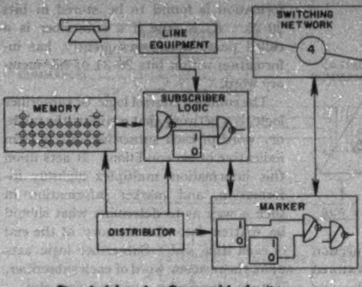


Fig. 1 (above). System block diagram

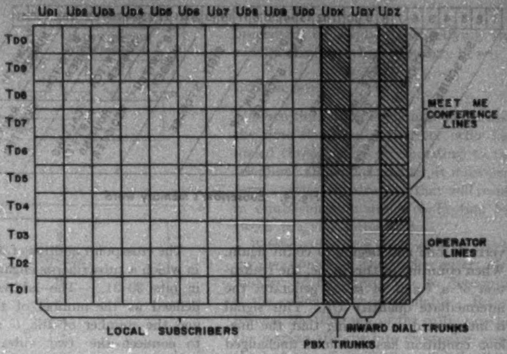


Fig. 2 (right). Time-slot assignments

seen by the subscriber logic, must remain constant for the duration of the time slot in order that all activities occuring within the time slot will be based on the same information. H_1 and H_2 are, therefore, sampled early in the time slot of the subscriber, and their sense retained within bistable multivibrators, or flip-flops, H_{S1} and H_{S2} for the duration of the reviewing period.

H_3, a third multiplex highway, will exhibit a binary "one" within the time slot of a subscriber designated to be the recipient of incoming PBX trunk calls during night-service operation. The signal emanates from the subscriber's line circuit, and its binary value is determined by a strapping option.

As has been noted, H_1, H_2, and H_3 are multiplex quantities. Each of the signals is generated within the line equipment of each subscriber, is associated with the distributor address of the subscriber, and is funneled into the subscriber logic via a logical network which is, in effect, a 130-input OR gate (Fig. 3).

Permanently associated with each time slot, and therefore with each subscriber, is a 31-bit memory-word that retains, between reviewing periods, pertinent information concerning the subscriber's activities; see Fig. 4. The memory word is read into 31 buffer flip-flops early in the subscriber's time slot; near the end of the time slot the information, as revised by subscriber logic, is written into the memory.

The first three bits of the memory word form a code which is descriptive of activity in which the subscriber is engaged. The eight resulting states are designated $S_1, S_2 \ldots S_8$, as subsequently listed:

S_1 = idle
S_2 = dialing first digit
S_3 = dialing second digit
S_4 = dialing third digit
S_5 = receiving busy tone
S_6 = connected—not talking
S_7 = connected—talking
S_8 = disconnecting

The next two bits form a second code, $Q_1, Q_2 \ldots Q_4$, indicating the relationship between subscriber and marker:

Q_1 = idle
Q_2 = demanding the marker
Q_3 = being serviced by marker as a calling party
Q_4 = selected by marker as a called party

A subscriber who is a called party and has dialed the number of a second subscriber is, if he is not connected to a trunk, attempting to establish a chain call. If, however, the subscriber is presently engaged in conversation over a trunk, the system must react by performing the operations necessary to effect a trunk transfer. It is apparent, then, that the system must be able to discern which subcribers are connected to trunks. Bit 6 of the memory word serves this end; a binary "one" is stored in bit 6 whenever a subscriber is involved with a trunk.

Memory bit 7 will contain a binary "one" if, during the previous reviewing period, information on the multiplex highways indicated that the subscriber's line loop was closed. By comparing bit 7 with the current status of the line loop, the transition from open to closed loop or from closed to open may be detected.

Bits 9, 10, and 11 form a counter capable of counting from zero to seven. Although the counter has been adapted to other uses, its primary function is to determine the length of time between changes in line-loop condition during states S_2, S_3, and S_4, in which dialing activity occurs. The counter begins to count when the first change in the loop condition is detected and advances one count every machine cycle that the line-loop status remains unaltered. When the line-loop condition changes before the counter reaches seven, the counter re-

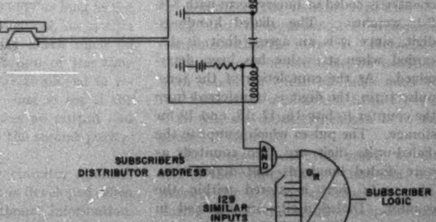


Fig. 3. Multiplex highway

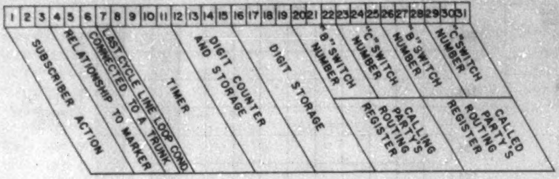


Fig. 4. Subscriber's memory word

verts to one and begins to count again. When counting in this mode, the realization of a count of seven generates the intermediate quantity Q_{S1}. This signal is interpreted as meaning that the line-loop condition has remained unchanged for a period of 16 milliseconds multiplied by seven, or 112 milliseconds. The maximum make-or-break period occurring during a dial pulse train is approximately 60 milliseconds. Therefore, Q_{S1} can never become true (be generated) until the pulse train is completed. When Q_{S1} is generated, assumptions may be made regarding the subscriber's action, based on the binary value of bit 7 at the time.

$P_7 . Q_{S1}$ = a dial-pulse train has been completed and subscriber is pursuing the call

$\overline{P}_7 . Q_{S1}$ = subscriber has hung up

A mark placed in bit 8 will make the counter of bits 9–11 perform a special counting operation. For instance, the system must be caused to disregard the dialing activity of a subscriber who has obtained a trunk and is outpulsing. The subscriber is given a limited time to accomplish this activity, and bit 8 regulates this relatively long counting operation.

Memory bits 12, 13, 14, and 15 are grouped to form a counter capable of enumerating line-loop interruptions generated by the subscriber's dial. The counter adds one each time a transition from open loop to closed loop occurs, and the count continues until Q_{S1} becomes true, denoting the end of dial-pulse train. The counter is coded in binary form with 8-4-2-1 weighting. The dialed hundreds-digit, since it is an access digit, is discarded when its value has been determined. At the completion of the tens-pulse train, the digit is transferred from the counter to bits 16, 17, 18, and 19 for storage. The pulses which comprise the dialed-units digit are then counted, as were dialed hundreds and dialed tens. The units digit is stored within the counter. Dialed digits are retained in their storage areas only for the duration of their usefulness.

The crosspoint address of a connection in which a subscriber is engaged is stored in bits 20–31. The address may be defined as the number of the *B* switch and the number of the *C* switch used to connect the two subscribers; see Fig. 5. (This is identical to the number of the junctor used to establish the connection.) The crosspoint address of a connection is stored within the calling party's register in bits 20–25. The first three bits of this group record the number of the *B* switch; the last three bits identify the *C* switch. A called party must refer to bits 26–28 and bits 29–31 to obtain the numbers of the *B* and *C* switches, respectively, through which his audio path is established.

Information within routing storage areas is written and erased under the influence of the marker during connection and disconnection cycles respectively.

Two important intermediate quantities are used extensively to describe a subscriber's status. L_S, which denotes a calling party, is generated when information is found to be stored in bits 20–25. R_S is true if a subscriber is a called party and, consequently, has information within bits 26–31 of his memory word.

The common control logic, or subscriber logic, is a network that converts the memory word of each subscriber into a form indicative of his condition. It acts upon this information, multiplex highway information, and marker information in such a way as to determine what should be written into the memory at the end of each time slot. Subscriber logic acts upon the memory word of each subscriber, in turn, during his assigned time slot.

The subscriber logic also exchanges information with the marker. To insure that the marker is not called upon to perform impossible or disallowed actions, the subscriber logic decides which subscriber demands are to be acted upon.

Certain supervisory signals (busy and dial tones) are supplied to subscribers via their line equipment. Subscriber logic regulates application of the signals.

Special subscribers—such as trunk and attendant lines—require special control and supervisory features not provided regular PABX subscribers. Inclusion of these features necessitates an expansion of the memory words of these lines, as well as the development of additional logic networks.

The Marker

The marker, in satisfying the demands of subscribers, may assume any one of 18

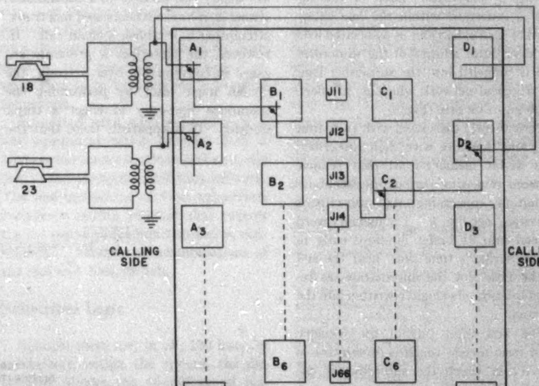


Fig. 5. Crosspoint-switching network showing subscriber 18 calling subscriber 23

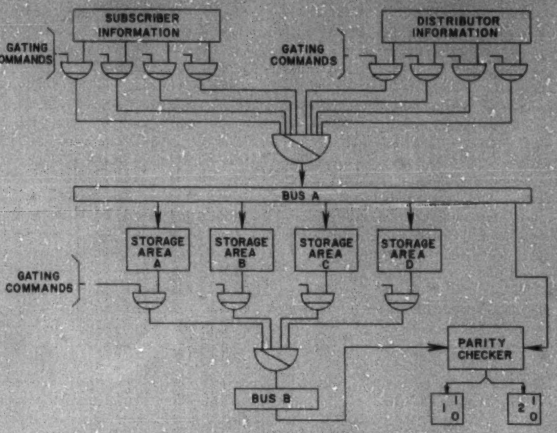


Fig. 6. Marker storage areas and parity checker

states. The state of the marker at a particular time determines its reaction to input signals. A number of flip-flops are used to record the state under whose influence the marker is acting.

It is essential that the framework of the marker include a number of storage areas, and that information be absorbed into these areas, processed, and discarded as a systematic response to predetermined conditions.

Fig. 6 depicts the form of the principal storage areas within the marker. Information from sources external to the marker is allowed to reach bus *A* when the explicit conditions of the gating commands are fulfilled. The information on bus *A* is made simultaneously available to the flip-flops that comprise storage areas *A*, *B*, *C*, and *D*, and it can be absorbed into any or all of the storage areas when expedient to retain it.

Just as information from external sources was made available to the marker by placing it on bus *A*, so information within storage areas *A*, *B*, *C*, and *D* is presented to the parity checker by gating it onto bus *B* at the proper time.

The parity checker is a logic network that compares two binary numbers, and yields an output signal whenever the input from two sources, buses *A* and *B*, is identical. The checker is useful in enabling the marker to seek out a particular time slot by one of two methods.

One method is based on the fact that a called party's number as dialed is identical to his distributor address. Consequently, if the dialed-tens and dialed-units digits, as recorded in the calling party's register, are stored in areas *B* and *C*, respectively, and are gated to the parity checker via bus *B* coincidently with the distributor-tens and units encountered during each time slot in turn, the time slot of the called party will be recognized by a parity-checker output, both while dialed-tens and distributor-tens and dialed-units and distributor-units are being compared.

The second method can find the time slot of one party to whom another is connected. Within each subscriber's register is recorded the crosspoint address of each connection in which the subscriber is currently participating. By storing in the marker the crosspoint address of a connection and then finding a time slot displaying an identical address, the party to whom another is connected may be located.

Because both methods of locating time slots depend upon the performance of two distinct parity checks during each time slot, and because the output of the parity checker persists only as long as the input signals remain identical, the results of an individual parity check must be preserved for the duration of the time slot. Two flip-flops are employed to retain parity decisions—one is set if the first evaluation produces an output and the other responds to the second parity-check.

Because the parity circuitry operates constantly, an output is developed when both buses are without information. Therefore, the output of the parity checker must be used judiciously. Indiscriminate setting of the flip-flops is prevented by explicitly specifying the conditions under which they may be set.

CONNECTIONS

When the marker is idle and encounters the time slot of a subscriber who has dialed three digits corresponding to the telephone number of another, it must perform a series of operations that will result in establishing an audio connection between the two subscribers.

Before concerning itself with establishing the connection, the marker enters a 16-millisecond busy-test cycle to determine if the called party is available. Before leaving the time slot of the calling party, the marker absorbs into its storage areas the last two digits he has dialed, which correspond to the distributor address or time slot of the called party. Through use of its parity-checking apparatus, the marker can identify the time slot of the called party and, by analyzing his memory word, determine if he is free to receive a call. No further action occurs until the calling party's time slot is again encountered.

With the question of subscriber availability resolved, the marker must next concern itself with selecting an available route through the crosspoint network to link the two subscribers. To understand this process, some knowledge of how the network is arranged is necessary. There are four stages of switching matrices or switches as seen in Fig. 5. Each subscriber is permanently connected to the *A* switch and to the *D* switch corresponding to his tens-group. Subscriber 49 would be represented on switches A_4 and D_4, for example.

Calling parties use their transformer connections to *A* switches; called party's audiopaths enter the switching network via *D* switches. Each *A* switch is linked to each *B* switch by one connection. Similarly, each *D* switch is connected to each *C* switch by one link. Each *B* switch is further connected to each *C*, switch via one junctor, a circuit which, under the control of the marker, regulates the firing and extinguishing of crosspoint elements within the network. Hence, for a connection to be available, for instance, from subscriber 34 to 92, links must be free from A_3 to a *B* switch, from D_9 to a *C* switch, and from the chosen *B* to the chosen *C* switch.

The route-search cycle starts at the end of the calling party's time slot, and persists for a complete machine cycle, during which the marker inspects the memory word of each subscriber within the system, and notes which *B* and *C* switches are in service. *B* switches currently in use by subscribers served by the same *A* switch as the calling party are listed by the marker as unavailable for the proposed connection, as are the *C* switches being used by subscribers of the same *D* switch as the called party.

During the cycle, the transmission network supplies the marker with information regarding junctor availability so that, when the time slot of the calling party is again reviewed, the marker has determined which, if any, routes may be employed in establishing the required connection and which one, if more than one are free, is to be used.

A successful route-search cycle is followed immediately by two complete machine cycles, during which the marker instructs the crosspoint switching network as to which route must be activated, and also tells the subscriber logic the nature of the connection so that memory-words of the participating parties may be revised accordingly. The marker then returns to the idle condition. The entire process of establishing a connection has taken a total of only 64 milliseconds.

THE HUNTING SEQUENCE

When called upon to connect a subscriber to any one of a group of lines (as in the case of a subscriber calling a trunk), the marker must select not only an idle member of the group but also one to which a free transmission path exists. Each type of line-hunting activity is characterized by a search for a line-fulfilling criterion, based upon some quality of the peculiar class of lines. A search for a trunk will culminate in the discovery of a time slot designated by U_{DX} or U_{DT}, because only trunks are assigned these locations. A time slot exhibiting H_8 would satisfy the marker when searching a line to which to connect an incoming PBX trunk during night-service hours. Of course, any line to be selected must be idle.

Another factor must also be considered. Once the marker has selected a particular line to which it will attempt to discover a free route, it must become insensitive to other lines which might satisfy the basic conditions of the search. When the marker enters the hunting state, a flip-flop, I_M, is set. The flip-flop remains set only so long as the marker is unable to locate a subscriber fulfilling the hunt criteria. When a line is selected, the flip-flop is reset; this prevents the marker from actively continuing the hunt.

After a line is selected as a prospective called party and its distributor address absorbed into the marker, the marker returns to the time slot of the calling subscriber and enters a route-search cycle. If successful, the marker instructs the transmission network to establish the connection. If unsuccessful and no free route is found, the marker re-enters the hunting state and attempts to select another line for use as a called party. In so doing, it must pass over the line previously selected. To accomplish this, every line, upon being selected as a called party, is acted upon to display state Q_4. The common criteria, which must be fulfilled in order for a line to be selected as a called party during a hunting action, have now been expanded to include:

1. The line must be idle.
2. The marker must not have previously selected a line during the same hunt sequence.
3. The line must not have been previously selected during the same call.

The use of the I_M flip-flop as the device that allows the marker to actively search proves advantageous; if no line is selected during the hunting cycle, I_M will still be true when the calling party's time slot is encountered—a condition that indicates to the subscriber that the desired connection is unobtainable.

TRANSFERS

The type of industrial or commercial concern most likely to require PABX facilities would be one in which areas of specialized activity are isolated from each other. Business calls over 2-way dial trunks or PBX trunks might well be of such nature as to dictate participation by representatives in more than one location. For this reason, transfer facilities were included in the PABX framework.

Any PABX subscriber may transfer the trunk to which he is connected to any other PABX subscriber within the exchange, or to the operator, simply by dialing the number of the second party. To explain the mechanics of establishing a transfer, a scheme of pictorial representation may prove helpful.

Figs. 7(A) and 7(B) show each subscriber's line equipment represented on both sides of the crosspoint network. Calls proceed from left to right, i.e., the calling party is considered to be on the left of the network and the called party on the right. Each subscriber may be connected as a calling party, a called party, a called party and a calling party simultaneously, or may be an idle party (no connections). No subscriber may act as the calling party for more than one connection at a time and no subscriber may act as the called party for more than one connection at a time. In other words a subscriber may have no more than one left-hand connection and one right-hand connection at any given time.

Assume that subscriber *A* has called one of the trunks *T*; see Fig. 7(C). A conversation is in progress and it becomes desirable to transfer the trunk to a second subscriber *B*. The solid line in the diagram indicates an existing connection; the broken line shows the proposed connection. For either *A* or *B* to remove themselves from the conversation without isolating the trunk, the new party *B* must be associated with the trunk and not with *A*. Effectively, the trunk will be the calling party although *A* will physically dial *B*'s number. In this case, the proposed connection may be readily established, since there is no connection to *T* on the left side of the matrix—i.e., *T* is not a calling party.

When *A* has completed dialing the number of *B*, the marker analyzes the situation and concludes that a transfer is required. It absorbs from *A*'s time slot the digits dialed and, as well, the transmission address describing the connection between *A* and the trunk. The marker reviews time slots until a memory word is found which exhibits a routing address identical to that of *A*. This will, of course, be the time slot of the trunk. Now that the marker is cognizant of the

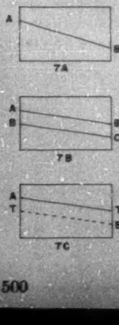


Fig. 7. Transfer action

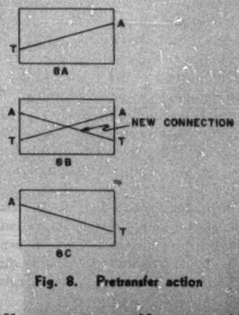


Fig. 8. Pretransfer action

distributor address of the trunk, the trunk may be treated as a calling party. The marker will perform the busy test, route search, and connection activities performed in the establishment of normal calls.

Assume, now, that conditions exist as depicted in Fig. 8(A). In this case, the trunk has called subscriber *A*. If transfer action is desired, difficulties are evident. The second local subscriber *B*, to whom the trunk is to be transferred, must be connected directly to the trunk as a called party. But the trunk, in this case, has already been connected as a calling party. Therefore, the immediate problem is to reverse the relationship between *A* and the trunk, causing *A* to become the calling and *T* the called party. The trunk can then act as a calling party in establishing the new connection to *B*.

Before the connection depicted in Fig. 8(A) can be destroyed, a new connection between *A* (calling party) and *T* (called party) must be established, as shown in Fig. 8(B), to insure that *A* and *T* remain continuously connected.

As soon as *A* demands the services of the marker, it recognizes that a pretransfer operation must take place. *A* will act as the calling party of a connection between himself and *T*. The only difference between the establishment of this connection and that of any other is that the busy-test sequence must be replaced by a cycle that will provide the marker with the distributor address of the trunk to whom *A* is connected. To achieve this end, the marker extracts the crosspoint-routing information which *A* and *T* have in common from *A*'s time slot, and reviews slots until the memory word of *T* is encountered. The marker then seizes *T*'s distributor address as being that of the called party.

Normal route-search and connection cycles follow. At the end of the connection cycle, the marker initiates a disconnection action to remove the connection between *T* and *A* in which *T* is the caller, see Fig. 8(C). Subscriber logic acts upon the memory word of *A* in such a way as to redemand the marker. This time, when the marker investigates the status of *A*, it will see fulfilled the conditions which require the normal transfer process.

Basically, the exchange utilizes the principle of calling party release—that is, only the party establishing a connection may cause that connection to be removed. The role of the marker during a disconnection action is two-fold: It must inform the crosspoint-switching network of the transmission path to be extinguished, and, during the time-slots of the participating parties, inform the subscriber logic of the action in order that memory words of the subscribers might be revised. The only information required by the marker to effect disconnection is the number of the route through the switching network that links the two subscribers, as recorded in each subscriber's memory word.

Conclusions

This type of telephone switching system was found to be quite flexible, and complicated features such as the transfer facility were readily incorporated into it. Perhaps the most difficult task was that of concisely defining the condition of the subscriber.

As in any logic network, redundant or "don't care" conditions were employed to reduce the network to a minimum. As a result, subsequent modifications and additions have been difficult to incorporate, and wasteful of logical elements.

The common-control concept creates reliability problems which are as yet unresolved

References

1. 100 Line Electronic Private Automatic Telephone Switchboard, John G. Van Bosse. *Automatic Electric Technical Journal*, Northlake, Ill., vol. 7, no. 2, Apr. 1960, pp. 50–57.

2. Features of an Electronic Crosspoint PABX, J. G. Van Vosse, J. DeCicco. *AIEE Transactions* (see pages 471–74 of this issue).

3. Transmission Network of an Electronic Crosspoint PABX, R. F. Kowalik. *Ibid.* (see pages 491–96 of this issue).

Reliable Data Transmission Through Noisy Media—A Systems Approach

C. M. MELAS
MEMBER AIEE

Synopsis: Error control techniques are becoming increasingly important in data transmission systems, since they allow reliable communications over media where the interference is high. Transmission errors can be corrected either by adding redundancy bits to the information to generate an error correcting code, or by automatically retransmitting the message when erroneously received. The relative merits of the two techniques are discussed in terms of both error control and systems compatibility. A class of powerful and easily implemented cyclic error detection and correction codes was recently developed which, in combination with automatic retransmission, should give effective error protection. Once the system reliability requirements are realistically gauged, and the characteristics of the channel interference are known, it is possible to determine a cyclic code best suited to that interference, and to determine a message length for the maximum transfer of information.

RELIABILITY requirements in information transmission range from the imperfect but intelligible transmission of a highly redundant signal, as a television recording, to the exact reproduction of messages of no inherent redundancy, as computer data. This paper is concerned with the latter case. Error control methods are considered which are designed to safeguard the high reliability of computer data throughout their transmission over relatively noisy channels.

Binary serial transmission over a single channel which is the simplest and most commonly used case will be considered. The original message is assumed to contain no redundancy; this is equally applicable to a message whose original redundancy has been eliminated by preprocessing such as Huffman coding. Redundant bits are introduced in the encoding process for the purpose of error detection and error correction. The final message is arranged in blocks; each block consists of several 6- to 8-bit characters.

A conventional error control system could operate by having redundancy bits included in the message (typically in the form of a parity check code) for the purpose of error detection. At the receiving end, the message is decoded in groups of binary digits; if an error is detected, transmission is interrupted with a demand for repetition of the last, erroneously re-

---

Paper 61-822, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted February 20, 1961; made available for printing April 24, 1961.

C. M. Melas is with the International Business Machines Corporation, San Jose, Calif.

The author is indebted to Dr. D. A. Linden and B. Hopner for their valuable suggestions and criticism during the preparation of this paper.

ceived, sequence. Contrast this system with one in which error control is achieved without a return channel by incorporating in the original message additional redundancy bits which enable the receiver not only to detect errors, but also to correct them. It is the purpose of this paper to compare these two methods, to indicate their domains of applicability, and finally, to suggest their simultaneous use in a system particularly well adapted to the stringent requirements of data transmission.

Essentially, error detection coding (EDC) and retransmission make an excellent system for reliable communication if the transmission channel introduces few errors, and if a high-quality return channel is available. The system deteriorates rapidly with increasing error rate. Error correction coding (ECC) is particularly effective if the characteristics of the channel are such that a relatively small percentage of the possible error patterns account for virtually all observed errors. If ECC is utilized to convert the actual channel into an effective channel having a relatively lower error rate, transmission may then be used to its best advantage. Although the principle of ECC has been known for years, it is only with the recent discovery of easily implemented codes that ECC has become attractive for practical application.

Error Detection by Coding

The simplest binary error detection code is the parity check, an additional digit chosen equal to the sum modulo 2 of all the bits in the message. The n-bit message, including the parity check, can be received in 2^n different ways, one of which corresponds to the original message. If an odd number of errors is present, the parity check will indicate an error. Thus, half of all possible errors can be detected with just one additional bit. The fraction of errors actually detected in transmission will depend on the characteristics of channel noise which assigns a given probability of occurrence to each of the 2^n possible versions of the message. Additions of a greater number of parity bits will increase the protection afforded by the code. In a message of arbitrary length including p check bits, the number of undetected errors will amount to a fraction $1/2^p$ of all possible errors, if the error detecting code is efficient, see Fig. 1. The undetected error rate can be made arbitrarily small by increasing p. The implementation of such codes requires only one element (flip-flop) per parity bit with very little added logic, this will be described under ECC Implementation.

The addition of a given number p of EDC bits detects a fraction $1/2^p$ of all possible errors regardless of the total length of the code word. As the word length increases, the frequently occurring error patterns form an ever smaller fraction of the total number of possible errors. Thus, the efficiency of EDC increases as the word length is increased, and a reasonably large structural subdivision of the message should be chosen as the code word length. Consider, for example, a block of 80 6-bit characters required to transmit the information contained in an 80-column punched card. A single parity check per character will not detect any double errors in a character, and will fail to recognize a double adjacent error, whose frequency in many channels is second only to single error occurrence. Double adjacent errors can be detected by the addition of a second check bit, equal to the sum modulo 2 of every other bit in the character. This checking scheme, however, will fail to recognize other common error patterns, such as a burst of four consecutive errors, and will add a total of 160 redundant bits, or 33% of the information bits. In contrast, in a block detection scheme, 9 parity bits would be sufficient to detect any double error in the message, or any burst up to 8 bits in length. The redundancy required will be less than 2%. Another method of error detection by character may be reached by fixed weight codes, where the information is encoded to include a fixed number of ones per character. An error will be detected whenever that number is altered by noise. Any odd number of bits in error will be found by this scheme as well as many even errors. An undetected error will occur when an even number of bits is corrupted in such a manner as to leave the total number of ones unchanged.

Fixed weight codes are attractive in unsymmetrical channels, where the probability of gaining ones by error is larger than that of gaining zeros, or vice-versa. A combination of fixed weight coding by character and parity checking by block has been successfully employed in some data transmission systems.

Error Correction by Coding

ECC, like EDC, involves additional check bits, which are the sum modulo 2 of certain information bits. The bit or bits in error can be pinpointed at the receiver by determining which parity bits fail to be equal to the sum of the corresponding information bits. In an n bit message including p parity bits, 2^p-1 erroneous versions of that message can be corrected, since there must be a one-to-one correspondence between every erroneous message to be corrected and every possible failure of the p sums. Of 2^n-1 possible erroneous versions of the transmitted message, the fraction $r=2^p-1/2^n-1$ can be successfully corrected with p check bits.

But,

$$r=\frac{2^p-1}{2^n-1}\approx\frac{1}{2^{n-p}}=\frac{1}{2^m}$$

where m is the number of information bits. The fraction of correctable errors is seen to depend only on the number of information digits, and to be independent of the number of parity digits; Fig. 2. The maximum fraction of correctable errors occurs for $m=1$ or for p parity bits and 1 information bit, where $r=1/2$. Therefore, with an error correction code consisting of separate information and parity bit groups it is impossible to correct more than half of all possible errors in a message, even though it may consist of just one information bit, and any number of check bits. By comparison, error detection of half of all possible errors can be accomplished with only one parity bit, regardless of the number of information bits as was described previously.

A system where each binary digit is repeated n times embodies the simplest example of an optimum error correction code; it consists of one information bit and $n-1$ parity bits. The decision at the decoder is based on a majority count of received bits. In such a scheme, it is possible to correct the received message only if half, or less, of the bits have been corrupted.

ECC is effective because not all errors in a message are equally probable. In some channels, the probability of certain

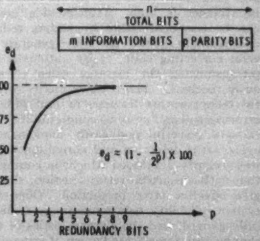


Fig. 1. Error detection efficiency

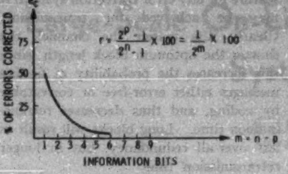


Fig. 2. Error correction efficiency

small classes of errors has been shown to be orders of magnitude larger than the probability of other classes. It is apparent, then, that an error correcting code can offer maximum protection only if it is designed specifically for the types of errors caused by the interference in the channel. These are generally a function of the type of modulation system employed. The distribution of the number of bits between successive errors, which will determine the optimum message lengths, and show whether errors occur independently or in bursts is particularly interesting. Error statistics of this type were obtained by Alexander, Gryb, and Nast for the United States switched telephone network.[1] They point out that errors are not independent but occur in bursts. The number of bits corrupted by a noise impulse is a function of the speed of transmission as well as the channel characteristics. This factor must be taken into account in determining a suitable error correction code.

ECC Implementation

It had been impractical to incorporate error correction in many systems because of the complexity of the decoding and correcting operations until recently. Abramson, who developed cyclic block codes, changed this situation.[2] The implementation of the burst error correcting codes developed by Peterson is not much more complex than that of cyclic error detecting codes.[3,4] The encoding operations for ECC and EDC are identical and require only a p-element feedback shift register, where p is the number of parity bits as in Fig. 3. The decoding operation is similar to the encoding operation, see Fig. 4. The feedback shift register is examined after information and parity bits have been passed through it, and if the storage elements are not all in the zero state, an error is present in the data. The error detection process ends there; a further cycle is necessary for error correction. With burst error correction, when a given bit configuration is tested in the feedback register, it is only necessary to add the contents on the feedback-register to the data being shifted bit by bit out of storage. The storage device need only be a serial buffer to delay the data, rather than a core buffer providing random access. This addition actually performs the correction of corrupted bits.

Cyclic codes have been developed both for the correction of burst and independent errors.[5–7] The same codes may be used for error detection only, or for a combination of error detection and correction. Since these codes all use the same implementation for error detection they impart a very large degree of flexibility to a system, and they may easily be modified to meet various reliability requirements.

The redundancy required for correction is not excessive. A code which corrects bursts up to 10 bits in length in the 480-bit block mentioned earlier requires 20 parity bits, or about 4% redundancy. For correction of any three errors occurring at random in the same block, 27 parity bits must be used for a redundancy of 5%.

An error correction code designed to correct a specific class of errors will interpret any error as belonging to that class. If the error is of a type that cannot be corrected, the correction process takes place anyway, and some good bits are altered. Furthermore, the alteration is made in such a way that it cannot be detected if the same code is used again.

Therefore, any system with ECC must have additional EDC performed to verify the correction cycle. Thus, three cycles comprise an ECC system:

1. An initial detection cycle where the parity relations are applied to the received data.
2. A correction cycle performed if any of the parity relations fail.
3. A final detection cycle to insure that no errors remain in the data.

Error Correction by Automatic Retransmission

The automatic retransmission of erroneous messages, where the receiver sends an acknowledgment to the transmitter at the end of every message is a most effective means of error control. When no acknowledgment is received, the transmitter then retransmits the same message.

In order to have a reliable automatic retransmission system, sometimes known as a feedback system, the following characteristics of the message are of importance:

1. The error detection applied to the transmitted message must be adequate since the whole system can be no more reliable than the error detection.
2. The acknowledgment signal must carry enough redundancy to be relatively immune to interference. Clearly, if the acknowledgment signal indicates a correct message, when actually the message has been detected as being incorrect, the transmission efficiency will be seriously impaired.

If a feedback signal indicates that the message is incorrect, when the message was correctly received, it will be retransmitted needlessly. This, however, will not impair the reliability of the transmission, but will result in decreased channel capacity.

An asymmetrical type of error correction could then be adopted for the acknowledgment signal. For example if k zero bits represent acknowledgment of a correct message, and k one bits are a repeat signal, the transmitter might repeat the message when m of the k bits it receives are ones and $m<k/2$. Clearly, the correction will be more asymmetrical as m is decreased, and the probability of needless retransmission will increase. The possibility of this latter occurrence makes it necessary to identify new or retransmitted messages at the receiving end by adding a group of bits at the beginning of the message in the form of an address code. The address code should also have enough redundancy to be self-correcting in order to protect against misinterpretation of a new message as a repeat. An excellent discussion of automatic retransmission is presented by Metzner and Morgan.[8]

With adequate redundancy, in the address at the beginning of each message and in the acknowledgment signal, re-

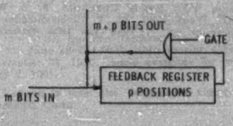


Fig. 3 (left). Encoding

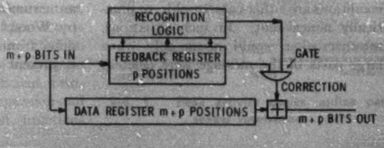


Fig. 4 (right). Decoding and correction

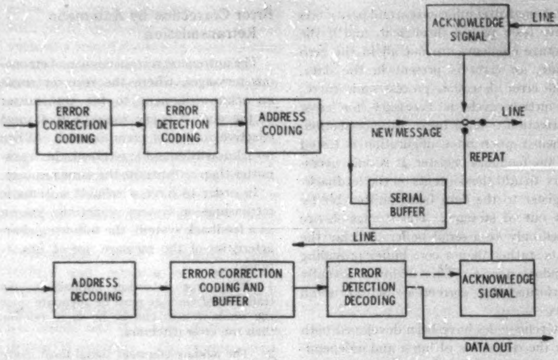


Fig. 5. Block diagram of transmission terminal combining retransmission and ECC

transmission is a very powerful scheme of error control, since it can be made as reliable as the error detection. Therefore, it is possible to attain an arbitrarily large degree of reliability, if channel capacity is not a factor. The main disadvantage of this method is that it requires 2-way communication in the form of a feedback channel which is not always available in transmission systems. Also, efficient use of the channel is not made because retransmission of a whole message regardless of the number of errors might be required. More redundancy than one-way transmission with error detection is needed because of the high reliability imposed on the acknowledgment and the address codes.

The following two examples show how the communication system may fail to meet the requirement of feedback transmission.

1. With radar data, or guidance systems, when transmission is in real time, automatic retransmission may not be possible since it would result in a variable time delay between the time the message is transmitted and the time it is correctly received, which is difficult to compensate for.

2. With simultaneous transmission of data from a master station to many different locations, automatic retransmission becomes extremely inefficient, since retransmission to all receiving locations is necessary if one of them sends a repeat request.

Automatic retransmission is particularly well suited to channels where noise conditions are either catastrophic, or practically nonexistent. In such situations, retransmission would occur only when large parts of the message were deleted. Thus, efficient use would be made of the channel.

When channel noise is such that only a small percentage of transmitted blocks are corrupted, over-all retransmission efficiency is high and any redundancy and hardware created by the addition of ECC would not be justified. This was the case in many wire line transmission tests where the fraction of blocks to be retransmitted was in the order of 2% or less.

Automatic retransmission per se is inefficient in types of channels where noise is sporadic. When a noise pulse occurs in every consecutive record, automatic retransmission alone is totally inefficient. Error correction coding, however, is useless in catastrophic conditions, but very effective for sporadic impulse noise. It may be achieved with little additional equipment since a buffer is available at the receiver for retransmission; see Fig. 5. A combination of the two methods would result in a very powerful tool for error control.

Optimum Message Length

Perhaps the most important factor in good channel utilization is an accurate determination of the optimum length of a transmitted message. This is true whether error correction is by coding or by retransmission. The block length can be determined once the error distribution of the channel is known, or calculated from an assumed noise distribution, as shown by Wood.[9] This message length might differ from other systems, but if a buffer is present at the terminal (a necessary feature for any error correction system) it may be achieved for transmission. Clearly, the noisier the channel, the shorter the optimum block length, since this increases the probability of finding messages either error-free or correctable by coding, and thus decreases retransmission time. Long blocks will result in less over-all redundancy, but in longer retransmission time.

The optimum block length for ECC need not be the same as for automatic retransmission, and generally will be shorter. This can be taken into account in a system where acknowledgment occurs only after several blocks have been received and where only the blocks in error are repeated.

Conclusions

Transmission through noisy media can be made arbitrarily reliable with the use of coding and feedback correction. ECC alone in its present form, cannot match the reliability obtained by retransmission, but it can effectively complement the latter by enhancing channel capacity. In systems where retransmission cannot be used, efficient and reliable transmission is possible by judicious choice of coding and block length if an accurate knowledge of error distribution is obtained through exhaustive testing.

References

1. Capabilities of the Telephone Network for Data Transmission, **A. Alexander, R. Gryb, D. Nast.** *Bell System Technical Journal*, New York, N. Y., May 1960.

2. A Class of Systematic Codes for Non-Independent Errors, **N. M. Abramson.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-5, 1959, p. 150.

3. Error Correcting Codes (book), **W. W. Peterson.** John Wiley & Sons, Inc., New York, N. Y., 1961.

4. Cyclic Codes for Error Detection, **W. W. Peterson, D. T. Brown.** *Proceedings*, Institute of Radio Engineers, Jan. 1961, p. 228.

5. A Class of Multiple-Error Correcting Binary Codes for Non-Independent Errors, **P. Fire.** *RSL-E2*, Sylvania Reconnaissance Laboratory, Mountain View, Calif., Mar. 1959.

6. A New Group of Codes for Correction of Dependent Errors in Data Transmission, **C. M. Melas.** *IBM Journal of Research and Development*, New York, N. Y., Jan. 1960.

7. On a Class of Error-Correcting Binary Group Codes, **R. C. Bose, P. K. Ray-Chaudhuri.** *Information and Control*, New York, N. Y., vol. 3, no. 1, Mar. 1960, pp. 68–79.

8. Coded Feedback Communication Systems, **J. J. Metzner, K. C. Morgan.** *Proceedings*, National Electronics Conference, 1960.

9. Optimum Block Length for Data Transmission with Error Checking, **F. B. Wood.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 855–61.

Large-Scale On-Line Data Processing Systems

S. LEVINE
MEMBER AIEE

ON-LINE DATA processing systems have been developed for industrial and business problems which may be classed as inventory control of products with regional, nation-wide, or even world-wide distribution and demand. The control and reservation of passenger or cargo space in the transportation industry, the handling of deposits and withdrawals in savings banks with many branches, the processing, distributing, and retrieval of information concerning stock transactions and quotations, and the control of inventory for large-scale manufacturing and distributing organizations as well as for government and military organizations are examples of application for which systems of this type are required.

All of these systems exhibit certain common characteristics which determine the requirements for an on-line system. These are:

1. A large number of geographically dispersed customers or subscribers who compete for access to the processing equipment and inventory can develop queries during heavy load periods.
2. The need for very fast response times to queries by agents' or subscribers' operating input devices, to reduce the number of agents and input devices required in peak periods and to minimize the time required to process a customer, who may be standing in line in front of a counter or teller's window, or connected by phone to the agent.

System Requirements

Experience in the design and analysis of on-line systems with the general characteristics listed previously has emphasized the following important requirements. These are particularly applicable to the airline transportation industry.

1. In addition to the need for fast response and access to input-output devices on a wide geographic distribution, the system must be designed to operate on a continuous basis (24 hours a day) with reliability and capacity to handle peak traffic.
2. The system must include facilities to assure accuracy of data in the system and the proper audit of all actions taken.
3. The system must have inherent capability of expansion to meet the expanding volume of traffic for at least 5 to 10 years after it is placed in operation.
4. The system must be flexible to permit changes in program as a result of changes in operating procedures.
5. The system must be flexible and expandable to permit integration of additional functions, which may not yet be fully defined, by means of added equipment or program changes as required. It must permit additions of this nature to be made without shutdown of those portions of the system required to maintain the traffic load.
6. The system must be designed so that a portion of it may be taken out of service for maintenance without upsetting the capacity to handle normal load.

System Design Approach

In earlier designs the equipment and systems design was tailored to one or two aspects of the problem for which specific requirements had been defined. Special-purpose wired program computers were used and extended into complete integrated systems by means of special communications systems.[1]

In considering the criteria outlined previously it has become apparent that the basic philosophy in approaching the design of on-line systems can no longer be based on the study and solution of only those aspects of the problem which appear to have the most immediate and urgent justification. An analysis should be made of all possible aspects of the problem on an integrated system basis.

The requirements and interrelationships of all functions should be defined, at least in general terms, and from this those areas which justify most immediate mechanization, as a part of an over-all plan of evolution into a complete system, can be established. This approach is considered necessary to assure that in proceeding with the design of systems to solve specific problems, sufficient flexibility is included to permit expansion and addition of equipment and functions without rapid obsolescence of the initial equipment.

Design Considerations

The airline reservation function imposes a peak demand requirement on the system. Fig. 1 illustrates the variations in traffic for a typical airline over a 24-hour period. This is shown both for an average day and a peak day. From this diagram it is clear that if one is to maintain rapid response times at all times, the system must be designed for the peak load requirements. This peak represents for many airlines only 1% of the total traffic handled by the airlines in a year and may not occur for more than a total of 25 to 50 hours during the year. Further, within any 24-hour day, the peak capacity for that day is required for no more than 6 to 8 hours.

Response time, as well as traffic volume, plays an important part in establishing the peak requirement. The nature of reservations operation is such that queries and responses for availability data and inventory actions should be processed by the system in a system-wide average of 1 to 2 seconds, and transactions involving passenger records in an average of 5 to 10 seconds. These requirements are imposed on the system by the fact that during peak periods an agent must be able to process a telephone inquiry in less than 2 to 3 minutes. Since each telephone call results in an average of five processor calls, each of which requires setup time and conversation time with the customer, a response time per call longer than an average of 5 to 10 seconds would affect the efficiency of an agent adversely. Thus, a system designed for this peak on-line requirement can have inherent capacity to handle all related off-line processing and updating requirements, as well as many other functions. A system designed to handle peak requirements with high reliability will be rela-

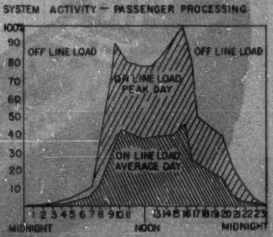


Fig. 1. Peak activity curves, airline passenger reservations

Paper 61-732, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted November 22, 1960; made available for printing May 3, 1961.

S. Levine is with the Teleregister Corporation, Stamford, Conn.

tively high in cost, hence economic justification may force inclusion of many related or even unrelated "off-line" requirements. ("Off-line" functions are defined as those which do not require immediate response.)

Traffic Analysis

Analysis of requirements for an airline passenger control system can be shown as a pattern of traffic generated at large numbers of offices (reservations offices, city ticket offices, and airport ticket offices). For a typical large nation-wide trunk airline, there may be as many as 100 to 200 offices in 50 to 70 cities. The number of input-output devices (agent or subscriber inquiry sets) required for a complete system may be 1,000 to 2,000, with quantities increasing because of increasing traffic at rates of 5–10% per year. The total traffic for an entire system, if handled by a central processor, ranges from five to 50 calls per second during peak periods, with messages averaging 100 to 400 bits. The range of message rates is a function of the size of the airline and the year considered as well as the functions included. A total processing load of this type requires a wide range of processor capability and justifies the use of a range of data communication channels ranging from 75 bits per second to 1,000 or 2,000 bits per second.

Response Time

The response time requirements for a system of this type are determined by the waiting time which can be tolerated by an agent and by the customer during peak periods. For an efficient and economical balance between equipment cost, agent cost, and customer satisfaction, response time for an airline inquiry should be from 1 to 5 seconds (measured from the time the agent requests access to the communication system and processor to the time the response starts to be received on his output display).

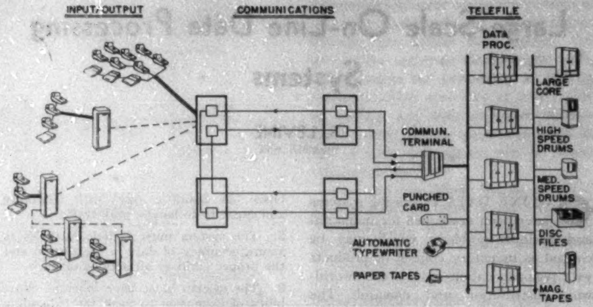


Fig. 2. Over-all system block diagram

Centralization Versus Decentralization

In considering the design of a nation-wide system, the problem of centralization or decentralization must be considered. If a system has traffic uniformly distributed or distributed in a number of concentrated areas, a decentralized system may appear to be justified. This approach would require a number of data processing centers with relatively short communications lines radiating as tributaries from each processor site. The problem of integrating these processors introduces the need for high-speed tie lines to interconnect the processors with a central processor.

Analysis has shown that, from a general communications point of view, the centralized approach is the more economical and provides maximum control. Each agent set acts independently, but control is centralized. Since the traffic and response time requirements of the airlines are approaching levels which require high-speed lines (1,000 bits per second) the centralized approach can achieve the tight control required.

Over-all System Design

A systems approach and design techniques have been developed by the Tele-

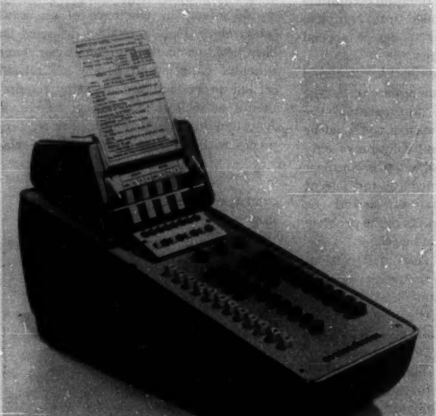


Fig. 3 (left). Typical airline reservation, input keyboard, display, and coded card

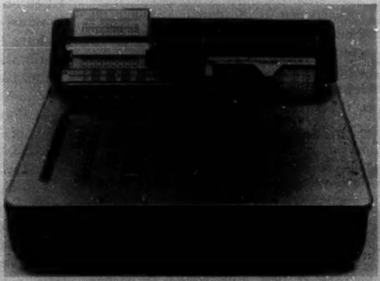


Fig. 4. Airline input unit with associated card printer in common housing

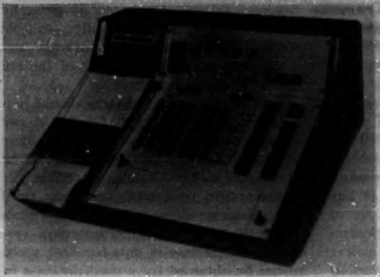


Fig. 5. Bank teller, input keyboard, and printer

register Corporation to meet the design criteria outlined above. The design is based upon a building block approach, which is expected to permit growth and expansion of the system in an economical fashion to take care of the growth of business, to permit addition of new devices for the input and output of data, and to extend storage and processing capability as needs develop. The building block approach permits installation of only those portions of the system required initially and expansion at the rate actually experienced, rather than requiring that the capacity for the full life of the system be provided initially. The system includes all elements of an integrated processing system, see Fig. 2, including input-output devices of various types, data communication links, and a stored program data processing system called the Telefile* Processor.

INPUT-OUTPUT DEVICES

A universal input-output device is difficult to achieve economically. Since, in an on-line system, many of these devices usually will be required, a special design can be justified. In general, the requirements are that each type of input-output device must be simple to use, must include a method of input checking, and should require a minimum of setup time to enter the data into the system. Some flexibility in design must be included, such as space for additional keys in a keyboard entry device,s pace for additional columns, or variability in letters or numerals in each column in a printing device.

A measure of modular design can be achieved even in the input-output position. This is achieved by design of each functional assembly as a compact subassembly which can be installed and combined with other assemblies in a common plastic housing. When made in medium to large quantities, these housings can be made cheap enough to be discarded when an additional device is added to the input position. Typical subassemblies can be a keyboard numeric array, a card reader, a line printer module, a page printer module, or an alphanumeric or alphabetic keyboard. Examples of arrangements of this type are illustrated in Figs. 3, 4, and 5.

Input-output devices in an on-line system are often grouped together in common locations where traffic requirements are heavy. For economy in use of communications lines, equipment is provided to interconnect and allot access to each agent and to provide necessary buffering and decoding to match the input-output device with the communications system.

COMMUNICATIONS SYSTEM DESIGN

The communications system design (Fig. 6) is based upon a flexible equipment configuration to provide the capacity, response time, and fallback justified for each reservations office. The large traffic generating centers are connected by high-speed (1,000 bits per second) full-duplex voice-band data channels to the data processor. Typical communications system geography is illustrated in Fig. 7. A number of these large remote centers called concentrators are connected in a single high-speed loop to minimize cost. A full duplex loop arrangement is employed in order to provide the reliability required for these heavy traffic centers. The loop arrangement provides two paths for reaching the processor and includes signaling facilities for monitoring the line for line or equipment failures, and for centralized control of the loop arrangement to bypass or isolate line sections in which the equipment or line has failed or become excessively noisy; see Fig. 8. More details of this equipment are discussed in another paper.[3] Further reliability can be achieved by

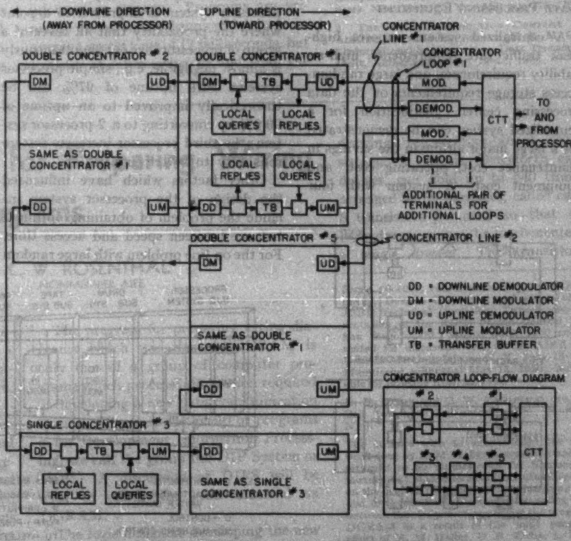

DOWNLINE DIRECTION (AWAY FROM PROCESSOR)
UPLINE DIRECTION (TOWARD PROCESSOR)
CONCENTRATOR LINE #1
CONCENTRATOR LOOP #1
DOUBLE CONCENTRATOR #2
DOUBLE CONCENTRATOR #1
DM
UD
TB
LOCAL QUERIES
LOCAL REPLIES
MOD.
DEMOD.
CTT
TO AND FROM PROCESSOR
SAME AS DOUBLE CONCENTRATOR #1
DD
UM
ADDITIONAL PAIR OF TERMINALS FOR ADDITIONAL LOOPS
DOUBLE CONCENTRATOR #5
CONCENTRATOR LINE #2
DD = DOWNLINE DEMODULATOR
DM = DOWNLINE MODULATOR
UD = UPLINE DEMODULATOR
UM = UPLINE MODULATOR
TB = TRANSFER BUFFER
CONCENTRATOR LOOP-FLOW DIAGRAM
#1
#2
#3
#4
#5
SINGLE CONCENTRATOR #3
SAME AS SINGLE CONCENTRATOR #3
LINE LAYOUT WITH DOUBLE & SINGLE CONCENTRATOR LOOPS.

Fig. 6. Communications system block diagram

Trademark, The Teleregister Corporation.

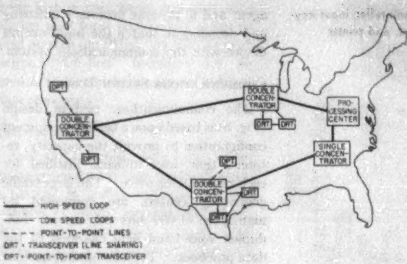

DPT
DOUBLE CONCEN-TRATOR
DPT
DOUBLE CONCEN-TRATOR
DRT
DRT
PRO-CESSING CENTER
SINGLE CONCEN-TRATOR
DPT
DOUBLE CONCEN-TRATOR
DRT
DRT
HIGH SPEED LOOP
LOW SPEED LOOPS
POINT-TO-POINT LINES
DRT = TRANSCEIVER (LINE SHARING)
DPT = POINT-TO-POINT TRANSCEIVER

Fig. 7. Communications system typical geography

redundancy in the concentrator stations. Dual units are provided with means for automatic switchover. The relatively fixed message lengths of the system permit use of the free-flow-type system. A transfer magnetic core buffer of capacity sufficient to accommodate the longest message is used at each concentrator to permit injection of single messages without endangering a message already in the line. This type of system permits fast responses or greater traffic-handling capacity than a selective-calling type of system, since capacity of the line is not affected by roll-call time. Expansion of the system is accomplished by adding stations or loops to the system.

Data Processing Equipment

A centralized system imposes high-peak traffic rate requirements, high-reliability requirements, and large random-access storage requirements on the data processing center. Justification for a centralized system can be demonstrated for a large major airline in the savings in maintenance cost, operating cost, and equipment cost. A system with four processor locations requires two processors at each site for reliability, and almost complete duplication of the storage at each location. A central system even with the same basic processors would only require four or five, and would provide better traffic-handling capacity and reliability. This is illustrated by queuing analysis. Multiple processors, when connected in tandem as multiple servers, can be loaded more effectively for the same delay or waiting time. A single processor can be loaded to 66% as compared to 88% per unit for a 3-processor system. Thus, a 3-processor system operated in parallel provides capacity equivalent to four separate processors.

Similarly, the reliability of a multiprocessor system is improved to a point where the probability that all three of a group of processors are out simultaneously is almost negligible, e.g., simple processor systems with uptime of 97% can be theoretically improved to an uptime of 99.9% by converting to a 2-processor system; the third processor increases system reliability to better than 99.99%.

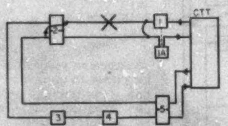
Other factors which have influenced the design of the processor system include the problem of obtaining optimum balance between speed and access time. For the on-line problem with large random access storage the system is generally access-limited with currently available storage techniques. The average on-line call uses more than 1/2 of the cell time for access to storage.

A 3-processor installation should provide adequate reliability for almost any application. The effective data-handling rate of the processor system should be designed so that two processors can handle the system peak traffic with the third processor serving as fallback and for other functions. With a multiprocessor system, an increase in data-handling speed can be obtained for a small increase in cost. This increase in processing rate need not be achieved by increasing the the basic speed of the processor. Reduction in effective access time to storage by use of parallel access paths can provide the improvement desired and avoid the more costly approach of a new or modified basic processor design for each application. Thus it appears that in considering all of the mentioned factors a system which can be set up in centralized multiprocessor arrangements represents the best approach to meeting the system criteria.

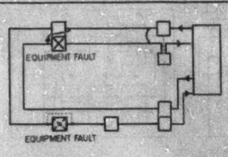
Processor Design

A new stored-program general-purpose data processor using solid-state components has been designed and constructed to meet these requirements. This Telefile Processor includes a number of unique features not found in processors currently in use; see Fig. 9. The central computer and control elements are organized around a relatively small magnetic-core working storage of 10,000 or 16,000 decimal digits. Operation of the computer is sequential by character at a rate of 16 microseconds per character.

Any single character in the core memory can be addressed and the word length is variable from one character to 100 characters or even up to the full size of the


CTT
2
1
1A
3
4
5

FALLBACK PROCEDURE - LINE OUTAGE


EQUIPMENT FAULT
EQUIPMENT FAULT

FALLBACK PROCEDURE—EQUIPMENT OUTAGE

Fig. 8. Communications fallback procedures

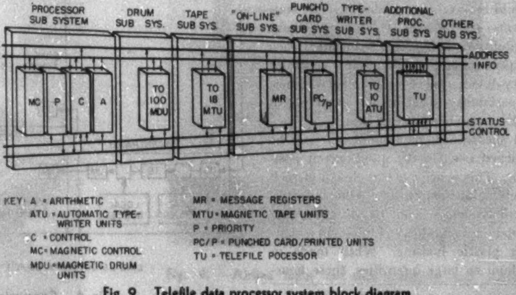

PROCESSOR SUB SYSTEM
DRUM SUB SYS.
TAPE SUB SYS.
"ON-LINE" SUB SYS.
PUNCH'D CARD SUB SYS.
TYPE-WRITER SUB SYS.
ADDITIONAL PROC. SUB SYS.
OTHER SUB SYS.
ADDRESS
INFO
MC
P
C
A
TO 100 MDU
TO 18 MTU
MR
PC/P
TO 10 ATU
TU
STATUS
CONTROL
KEY: A = ARITHMETIC
ATU = AUTOMATIC TYPE-WRITER UNITS
C = CONTROL
MC = MAGNETIC CONTROL
MDU = MAGNETIC DRUM UNITS
MR = MESSAGE REGISTERS
MTU = MAGNETIC TAPE UNITS
P = PRIORITY
PC/P = PUNCHED CARD/PRINTED UNITS
TU = TELEFILE POCESSOR

Fig. 9. Telefile data processor system block diagram

core memory by use of end sentinels. The central processor also serves as a control and switching center to permit communication over a common 4-bus network with many different peripheral subsystems. The basic structure permits interconnection to up to 29 subsystems, each of which can operate independently! Communication to and from the processor is controlled by a priority control system.

This structure permits multiplexed operation and overlapping of computer internal operations with subsystem operation such as: location and selection of information on magnetic drums and tapes for subsequent transfer to and from the central working memory, or transfer of a single character every 100 milliseconds from a number of slow-speed input-output subsystems such as paper tape punches and electric typewriters.

Communications to remote on-line input-output devices over communication channels is also accomplished through one or more peripheral subsystem channels through buffer-register subsystems. Subsystems such as magnetic drums, magnetic tapes, magnetic discs, electric typewriters, and tape punches and readers, do not include buffer registers but by means of a small control register can address and transfer information to and from sections of the core memory.

Further expansion of the Telefile Processor System can be accomplished by additional modules which permit two processors to be operated in either a dual or duplex mode. In dual mode, both processors operate simultaneously on the same data fed from a common input-output subsystem. Dual operation permits cross check of each operation. In a duplex or multiplex mode, each processor operates independently but can have access to a common pool of peripheral subsystems through a switching or selector unit. This facility is essential for on-line systems to permit parallel operation of processors on messages arriving over a common high-speed communications link and to permit access of any processor to all information stored in the main memory magnetic drums. An interlock system is provided to exclude a second processor from access to a particular section of the drum memory, or to a particular magnetic- or punched-tape unit until the first processor has completed its operation and released the subsystem.

Some queuing interference does occur, but it is relatively small because in on-line operation a large portion of the operations are inquiries to information stored on magnetic drums, and only one access is required to the information, after which it is transferred to the core memory for subsequent operations.

Through a single drum-control unit, each processor can have access to many large drums, each of which can store one million decimal digits with complete single-digit addressibility. Thus, the drum serves as an extension of the central core memory for program as well as information storage. Programs are primarily stored in the drums which have an average access time of 17 milliseconds, but transfer of information to and from the drum is controlled at clock rate after the data to be transferred is located. A few frequently used programs are retained in the core memory, as well as control information to locate the programs on the auxiliary drums.

Other peripheral subsystems not yet designed can be added to the system, as required, after installation. This could include new random-access memory devices, or a separate large auxiliary magnetic-core memory for use in those problems which require faster access to programs.

Conclusion

1. Large-scale, integrated, on-line data processing systems are economically and technically feasible today.
2. The design of such systems requires a user- and integrated system-oriented approach as opposed to the hardware-oriented approach.
3. Such systems, which represent advance in system design rather than any unique hardware development, are now in the process of cutover for a nation-wide airline and for several large regional savings banks.

References

1. Design Techniques for Multiple Interconnected On-Line Data Processors, F. J. Gaffney, S. Levine. "Proceedings of Eastern Computer Conference," Institute of Radio Engineers, New York, N. Y., 1957, pp. 173–177.

2. The Telecommunication Complex of An Airline Reservation System, H. K. M. Grosser. "International Symposium on Data Transmission," Delft, Netherlands, Sept. 19–21, 1960.

Computer Program for Preparing Wiring Diagrams

D. B. KIRBY
NONMEMBER AIEE

C. W. ROSENTHAL
NONMEMBER AIEE

THE PROGRAM under discussion prepares chassis wiring diagrams from a list of the chassis terminals to be connected. The diagrams are figurative, representing the existence, but not the exact location, of real wires on the chassis. These diagrams are a two-dimensional form of wiring information and are therefore of more value in maintenance and production than a single- or double-ended wiring list. In particular, the diagrams may be used to record changes, trace signals, and guide wiring shop personnel.

The program for preparing wiring diagrams which was named DGRAM is only one of a group of computer programs which process information required in developing a new data communication system. Because the group of programs is for Development Information Processing, we call the group the DIP System or DIPS. A discussion of DIPS will be used to orient the later material on DGRAM.

The logic designers developing the new system prepare their information as logic drawings representing about 200 transistors. A chassis, henceforth called a drawer, can hold up to 640 transistors arranged as 80 packages, called cards, of 8 transistors per card so that several logic drawings may be implemented with a single drawer. The transistors re-

---

Paper 60-1007, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, San Diego, Calif., August 8–12, 1960 and re-presented for discussion only at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted May 13, 1960; made available for printing April 25, 1961.

D. B. Kirby and C. W. Rosenthal are with the Bell Telephone Laboratories, Inc., Murray Hill, N. J.

The authors wish to acknowledge the extensive work done by N. M. Haller, a summer employee now returned to his studies with the Engineering Department at Massachusetts Institute of Technology, on the over-all planning and programming of DIPS, and to thank R. Neddermeyer for acting as liaison between the equipment people and the programmers. DGRAM is a result of the joint programming effort of N. M. Haller, D. B. Kirby, L. Boginsky, and A. W. Roberts.

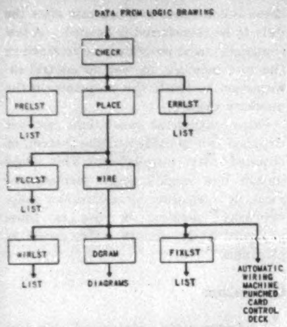

DATA FROM LOGIC DRAWING
CHECK
PRELST
PLACE
ERRLST
LIST
LIST
PLCLST
WIRE
LIST
WIRLST
DGRAM
FIXLST
LIST
DIAGRAMS
LIST
AUTOMATIC WIRING MACHINE PUNCHED CARD CONTROL DECK

Fig. 1. Dip system flow chart

quired by the logic drawings are assigned by an engineer to locations on cards. Then each card to be placed in a drawer is given a symbolic designation from 1 to 80. There is a prearranged correspondence between transistor locations on a card and the terminals utilized for access and egress of signals, grounds, and power voltages. Thus assignment of logic transistors to a card corresponds to designating the card terminals on which wires start and stop. However, the location of cards on a drawer has not yet taken place.

From this point on, DIPS is used to (1) optimally place cards on the drawer, (2) prepare instructions for an automatic wiring machine, (3) print wiring diagrams, and (4) print lists indicating the placement of cards, wires, terminal designations, etc., in the various formats needed. These tasks are accomplished in the sequence shown on the flow chart of Fig. 1. The flow chart names identify the particular programs referred to above. Note that a checking program precedes the others to weed out errors before substantial work is done and that a list of such errors is printed for the designers.

In operation, DIPS utilizes the units shown in Fig. 2. Data cards are keypunched and verified after tabulation of information from the logic drawings. The engineer controlling DIPS selects the programs and the exact sequence in which they are to be performed. He then lists the names of these programs on International Business Machines Corporation (IBM) cards which are called order cards. The data and order cards are combined into one deck per drawer. Three tapes are also required: one tape for the files of previously processed data (arranged as one file per chassis), one tape to receive the files of newly processed data, and one tape for the DIPS library of programs. DIPS contains a master program which

Fig. 2 (right). Dip system operational units

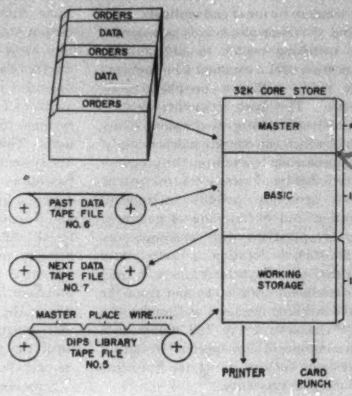
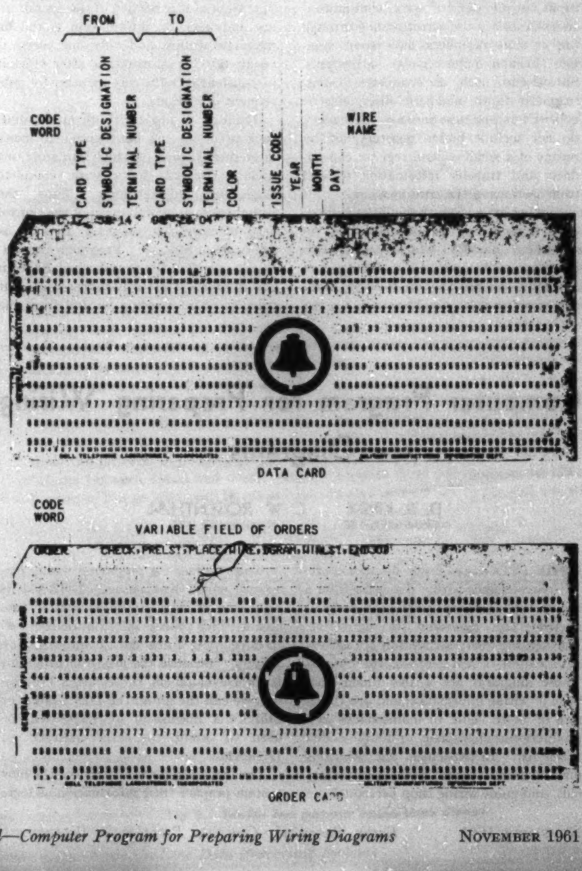

ORDERS
DATA
ORDERS
DATA
ORDERS
32K CORE STORE
MASTER
≈8K
BASIC
12K
WORKING STORAGE
12K
PAST DATA TAPE FILE NO. 6
NEXT DATA TAPE FILE NO. 7
MASTER PLACE WIRE.....
DIPS LIBRARY TAPE FILE NO.5
PRINTER
CARD PUNCH

Fig. 3 (below). Data card and order card


FROM
TO
CODE WORD
CARD TYPE
SYMBOLIC DESIGNATION
TERMINAL NUMBER
CARD TYPE
SYMBOLIC DESIGNATION
TERMINAL NUMBER
COLOR
ISSUE CODE
YEAR
MONTH
DAY
WIRE NAME
DATA CARD
CODE WORD
VARIABLE FIELD OF ORDERS
ORDER CHECK,PRELST,PLACE,WIRE,DGRAM,WIRLST,END
ORDER CARD

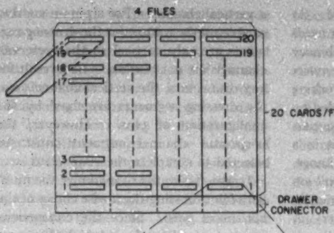


FIG. 4

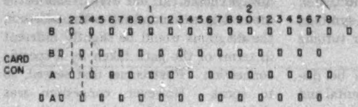


FIG. 5

Fig. 4 (top). Arrangement of cards in a drawer

Fig. 5 (bottom). Arrangement of terminal pins on a card

reads the order cards and directs control to the listed programs such as PLACE, WIRE, and PLCLST. Data and order cards are illustrated in Fig. 3 and their contents labeled.

The arrangement of the 32,000-word IBM *704* core store is shown in Fig. 2. There is allotment of areas for the master program of DIPS and for working storage. In addition, 12,000 words are set aside for a compressed representation of all information describing the connections on a drawer. Up to 3,000 interconnections are permitted and each requires four words. These four words hold the data introduced on the data cards plus that obtained as a result of the various programs. This area, called Basic, is "passed" from program to program, each extracting needed data and inserting results.

The programs may be run in two ways: (1) in the natural sequence of Fig. 1, or (2) individually. Engineering changes may be introduced and the entire sequence of programs rerun, or only portions may be rerun. Data in Basic may be retained on tape for future processing and as a record of the drawer. The wiring diagrams will be described following a review and detailed description of the equipment layout.

The 80 cards are plugged into the drawer in four columns or "files" of 20 cards each. So now one has one drawer, four files/drawer, 20 cards/file, as shown in Fig. 4. Each card contains 28 "terminals" and two "pins"/terminal which are electrically connected. In addition, two ground

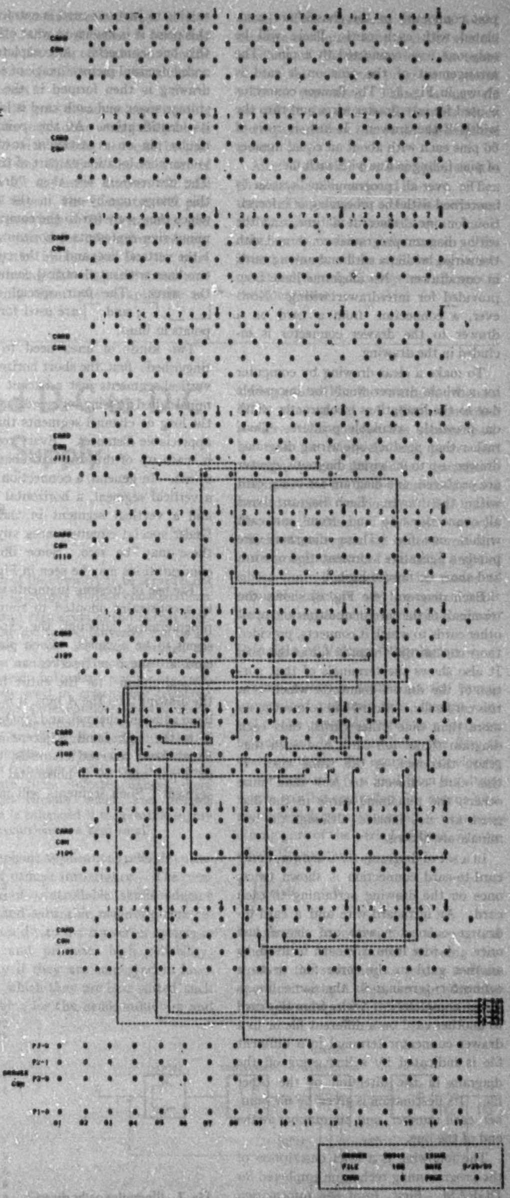


Fig. 6. Illustrative wiring diagram

pins connected to the chassis are associated with each card. There will be only one wire connected to a pin. The arrangement of the pins on a card is shown in Fig. 5. The drawer connector is used for interdrawer wires and runs the width of the drawer. It has six rows of 66 pins each with about an equal number of pins falling in line with each file.

The over-all programming system is concerned with the processing of information for one drawer at a time, and the wiring diagram program is concerned with the wiring within a card and among cards in one drawer. No diagrams have been provided for interdrawer wiring. However, a connection from a card on a drawer to the drawer connector is included in the drawing.

To make a clear drawing by computer for a whole drawer would be impossible due to the limitations of character width on presently available printers. Thus, rather than produce one wiring diagram/drawer, up to 80 wiring diagrams/drawer are produced, one diagram for every card within the drawer. Each diagram shows all connections to and from one card within one file. These diagrams are purely a figurative representation of wires and show no flow of logic.

Each diagram (see Fig. 6) shows the terminals of the card in question and of all other cards to which it connects, provided those cards are in the same file as this one. It also shows the terminals of that portion of the drawer connector which is in this card's file. A card can connect to no more than nine other cards, thus each diagram shows ten cards and each diagram, therefore, is the same size If this card connects to less than nine others, the remaining cards in the diagram are not labeled although the terminals are shown.

In a set of diagrams for a drawer, every card-to-card connection is shown twice, once on the drawing pertaining to each card. An intracard wire and a card-to-drawer connector wire are shown just once. A wire from this card to itself, to another card in the same file, or to a connector terminal in the same file, is shown completely. A wire from this card to another card on a different file or to a drawer connector terminal in a different file is indicated by a line going off the diagram in the direction of the other file. Its destination is given by file number, card number, and pin number at the end of the line.

The following is a brief description of the programming technique employed for DGRAM. First, from the information for the whole drawer, the information pertaining to just one card is extracted. At this point it is known to what other cards this one connects. A complete binary coded decimal picture image of the entire drawing is then formed in the working storage space and each card is labeled by its identification. At this point a particular pin on a particular card is in a known core location as part of the image. The connections are then "drawn" on this image one-by-one in the order in which they were fed to the computer. A minus sign represents a horizontal line, a 1 the vertical line, and + the crossing of two lines without electrical connection of the wires. The four special symbols, └, ┘, ┌, and ┐ are used for turning points in lines.

Two kinds of lines need to be distinguished: first, the short horizontal and vertical segments just adjacent to a terminal called dressing segments and second, the long or channel segments that travel appreciable distances. Every connection is made up of dressing and channel segments. In general, a connection requires a vertical segment, a horizontal segment and a vertical segment in that order. Under special circumstances any one of these may be zero. Some illustrative configurations may be seen in Fig. 7.

The use of dressing segments is related to a constraint adopted to simplify the programming, although the explanation seems to be complex. Each pin on the card in question reserves an adjacent vertical channel for the entire length of the drawing; for the *A* pins, it is the left-hand adjacent channel, and for the *B* pins, it is the right-hand adjacent channel. From these reserved channels, one gets to a pin only by a horizontal dressing segment. Such segments will not block a vertical channel. For a pin on another card or on the connector, the dressing segments are only vertical so that reserved channels are not blocked. As a result for any connection, the vertical channels and the dressing segments are fixed by the configuration of pins. However, the horizontal channel segment must be selected to complete the connection.

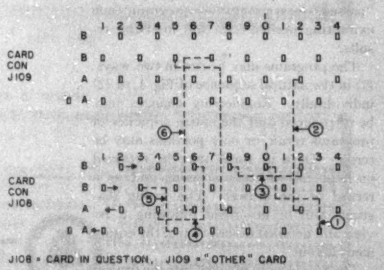
It was attempted to minimize the number of line crossings by a wise choice of the horizontal path. Since the connections are drawn in the order in which they were given the computer, the best path is not always chosen. If one were to shuffle the input data cards and rerun the program, the diagrams would be slightly different in terms of the path taken by a specific connection. First come, first served, so to speak. However, congestion was eliminated about the card in question by choosing the horizontal paths in the following ways:

For intracard wiring, the initial try for a horizontal path is one line below the left-most terminal.

For intercard wiring, the horizontal path is chosen one line below the terminal on the other card. This gets a good portion of the lines away from the card in question. This idea is also used in the card-to-connector case, except the horizontal is chosen above the connector terminal since the connector is at the bottom of the picture.

For those connections going to a connector pin on another file, the horizontal is always chosen between the first card and the connector. For those connections going to card pins on another file, the horizontal is chosen immediately below the pin on the card in question.

If the first choice for a horizontal is al-



	2 DRESSINGS	LEFT VERTICAL	HORIZONTAL	RIGHT VERTICAL
1	✓	✓	✓	✓
2	✓	✓		
3	✓		✓	
4	✓		✓	✓
5	✓		✓	✓
6	✓	✓	✓	

Fig. 7. Illustrative paths drawn on a card

ready occupied, further choices are made in a predetermined fashion depending on the type of connection. A specified number of tries are made and if the connection cannot be made, it is saved until all other connections on this card are made. At this point one looks for a vacant horizontal with no limits on the number of tries for a free channel. It is anticipated that this situation will occur most frequently on the intercard and card-to-connector types. In these cases the vertical channel to the other pin may be blocked on top or bottom by the presence of a second pin. In this case, the vertical channel adjacent to the pin on the other card or connector is tested. If it is free locally, this connection is permitted to enter on the side of that pin. In the runs made so far, however, this situation rarely occurs.

One of the greatest advantages in using a computer to produce wiring diagrams is the speed with which they can be produced. Using DIPS to obtain a set of drawings for an entire drawer will take from 2–5 days. The computer will produce the drawings in 12–15 minutes, and the other time is for tabulating and keypunching the information for the computer, printing the diagrams from tape and having them photographically reduced. This represents about a 5-to-1 decrease in elapsed time over the manual handling of this information. In terms of cost the computer program is about 1/10 as expensive. These savings do not reflect the additional benefits which are expected from the fewer anticipated errors.

After a drawing has been printed, it is photographically reduced in size. Then the set of drawings for a chassis are bound with a cover sheet containing an index, date, and revision information. This bound volume will be updated by subsequent individual drawings and changes in the cover sheet.

A Magnetically Regulated D-C to D-C Converter Power Supply

GUENTER SAGER
NONMEMBER AIEE

IN SITUATIONS where a single d-c source only is available or desirable to power electronic equipment requiring a number of different and isolated supply voltages, a so-called d-c to d-c converter[1–3] can be used. A typical example of this situation is found with mobile equipment where space and weight considerations determine the available power source. Communication equipment having buffer batteries to maintain continuous operation, which are unaffected by interruptions of the primary power source, is another example.

In the specific case of a power supply for the transistor circuits of an electronic telephone exchange, the d-c to d-c converter principle offers a number of advantages:

1. Existing battery installations can be used.

2. Since a d-c to d-c converter permits derivation of any number of different and isolated voltages from a single d-c source, the design of the circuitry to be served is not restricted by supply voltage considerations.

3. Efficient regulation of the output voltages of the d-c to d-c converter can be accomplished in the a-c stage through a-c regulation methods. The terminal voltage of lead-acid batteries, which are predominantly used in present-day telephone installations, varies considerably with charging and loading conditions and with ambient temperature. Changes of as much as ±20% from a nominal voltage are encountered. Although transistor circuits can be designed to tolerate these voltage changes, more economical circuits result and reliable operation is enhanced if the possible supply voltage excursions are kept small.

In telephone applications reliable operation is of utmost importance. The combination of controllable semiconductor devices and saturable magnetic core reactors in a d-c to d-c converter system is efficient and promises high reliability, especially if they are employed in functions for which they are best suited, such as switching for the semiconductors and timing control for the saturable reactors. With the advent of silicon-controlled rectifiers,[4] this concept has become feasible even for relatively high power and voltage levels. Regulated transistor–magnetic-core d-c to d-c converters have been described previously.[5] This paper discusses the design of a 24-volt converter which effects regulation of its 9-volt output through control of the firing angle of silicon-controlled rectifiers by means of saturable reactors. A low-power 2-volt output is regulated by a transistor circuit.

Principle of Voltage Regulation

Since power is defined as the energy transmitted per unit time, lossless control of the power delivered to a load can be achieved by constraining the flow of energy in time. For example, the electric power delivered to a load from a voltage source can be controlled by means of a switch which permits current to flow only during part of the period of time, T, over which the average

$$\frac{1}{T}\int_0^T ei\,dt$$

is formed. In contrast to the method where the excess power is being dissipated, the switching or gating method operates without loss, except for the losses introduced by the shortcomings of

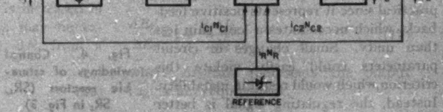


Fig. 1. System block diagram

Paper 61-150, recommended by the AIEE Magnetic Amplifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted October 28, 1960; made available for printing April 6, 1961.

GUENTER SAGER is with General Dynamics/Electronics, a Division of General Dynamics Corporation, Rochester, N. Y.

The author would like to thank J. D. Ross, Jr., and J. A. Tabaczynski, who contributed experimentation and discussion during various phases of the developmental work leading to this paper.

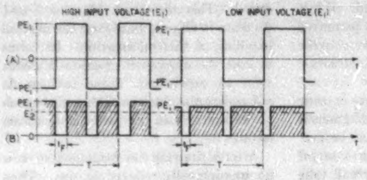


Fig. 2 (left). Voltage waveforms illustrating regulation concept

A—Input to control and rectification unit (point a in Fig. 1). The general case where the frequency of the square wave is proportional to the input voltage ($E_1T/2=\text{const}$) is shown

B—Output of control and rectification unit (point b in Fig. 1) maintaining constant average voltage E'_2 (shaded area) by adjusting prefiring interval t_F

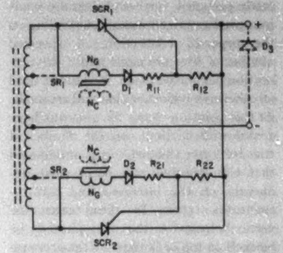


Fig. 3. Center-tapped full-wave magnetic amplifier with controlled rectifiers

a practical switch. The switching principle can be applied to regulate the output voltage of a d-c to d-c converter power supply against variations of input voltage and output current.

Fig. 1 depicts the over-all concept as a block diagram. The input direct voltage, E_1, is converted into a symmetrical square wave by the chopper. After being transformed to the desired level by the transformer of ratio $1:p$, the square wave is fed into the control and rectification circuit. The output of this circuit, whose waveform is shown in Fig. 2(B), resembles a pulse train of twice the frequency of the square wave (full-wave rectification). The pulse width is controlled during rectification in such a way that the volt-seconds per unit time supplied to the output remain constant. The output filter, acting essentially as an "averager," smoothes the pulses into direct voltage E_2. In principle, it is irrelevant where the pulse width control takes place; it could be anywhere between the chopper and the output filter. However, because of the limitations and shortcomings of actual devices, the practical solution consists of a circuit which essentially operates like a center-tapped full-wave magnetic amplifier, except that the power is gated by a pair of silicon-controlled rectifiers (SCR's); see Fig. 3. This combines the functions of control and rectification into one block as shown in Fig. 1. The prefiring interval, t_F, is controlled by the ampere-turns, $i_{1c}N_{c1}$, to stabilize the output voltage against variations of the input voltage, and by $i_{c2}N_{c2}$ to regulate for the output voltage drop caused by the output current, I_2. Making i_{c2} directly proportional to I_2 is not practical since it represents positive feedback, which necessitates a loop gain less than unity. Small changes in circuit parameters could easily violate this criterion, which would result in instability. Instead, the regulating signal is better obtained as the difference between the output voltage and a reference voltage, or, as shown in Fig. 4, as the difference, d, of the ampere-turns, $i_{c2}N_{c2}$ and i_RN_R, derived from these voltages. This type of regulation represents negative feedback. The amount of permissible loop gain is restricted by the response of the loop comprising the magnetic amplifier and the filter. The regulation against changes of input voltage E_1 is uncritical. No stability problem exists when i_{c1} is made directly proportional to E_1, since it is an open-loop control.

For the purpose of the following analysis, the drop of the output voltage with increasing output current is described in terms of a lumped internal resistance, R_i, in series with the output lead.

$$E_2'=E_2+R_iT_2 \tag{1}$$

Since the square wave is symmetrical, it suffices to integrate waveform (B) of Fig. 2 over only one-half of period T.

$$E_2'=\frac{pE_1\,T/2-pE_1\,t_F}{T/2} \tag{2}$$

pE_1t_F represents the amount of volt-seconds absorbed by one saturable reactor of the magnetic amplifier. It is a function of $\Sigma(i_cN_c)$, the sum of the control ampere-turns acting on the core. As the number of turns of the gate winding is N_G, we can write

$$pE_1\,t_F=N_G+f\Sigma(i_cN_c) \tag{3}$$

In this way $f\Sigma(i_cN_c)$ is defined as the control characteristic of the saturable core. Substituting into equation 2, the output voltage becomes

$$E_2'=pE_1-N_G\,\frac{f\Sigma(i_cN_c)}{T/2} \tag{4}$$

The control characteristic, $f\Sigma(i_cN_c)$, an intrinsic property of the magnetic core, cannot be expressed easily in analytical form. Therefore, a method will be developed which determines the various design parameters of the regulator system graphically. Essentially, it consists of plotting equation 4 versus $i_{c1}N_{c1}$, where E_1 is expressed in terms of $i_{c1}N_{c1}$, while defining

$$\Sigma(i_cN_c)=i_{c1}N_{c1}+d \tag{5}$$

As already mentioned above, it is

$$d=i_{c2}N_{c2}-i_RN_R \tag{6}$$

A set of curves is drawn for a number of arbitrary values of parameter d permitting the selection of suitable values of d to fit equation 1.

Two practical cases have to be distinguished:

1. The chopper frequency is constant: $T/2=\text{const}$.
2. The chopper frequency is not constant. It is proportional to E_1, if the chopper is

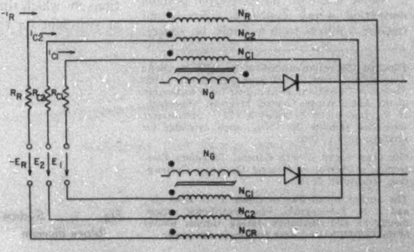


Fig. 4. Control windings of saturable reactors (SR_1, SR_2 in Fig. 3)

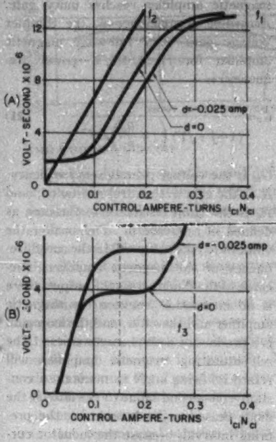


Fig. 5. Graphic representation of equation 9

$$f_1 = f(i_{c1}N_{c1}+d)$$

$$f_2 = \frac{pTR_{c2}}{2N_GN_{c1}} i_{c1}N_{c1}$$

$$f_3 = E_2{}' \frac{T}{2N_G}$$

controlled by a saturable core transformer (T_1 in Fig. 7), making $E_1\ T/2 = \text{const}$.

CASE 1

Multiplying equation 4 by $T/2N_G$

$$E_2{}'\frac{T}{2N_G} = pE_1\frac{T}{2N_G} - f\Sigma(i_cN_c) \tag{7}$$

If we define

$$i_{c1} = \frac{E_1}{R_{c1}} \tag{8}$$

as indicated in Fig. 4, equation 7 becomes with equation 5

$$E_2{}'\frac{T}{2N_G} = \frac{pTR_{c1}}{2N_GN_{c1}} i_{c1}N_{c1} - f(i_{c1}N_{c1}+d) \tag{9}$$

A graphic method for finding a range of $i_{c1}N_{c1}$ for which

$$E_2{}'\frac{T}{2N_G} \approx \text{const}$$

is shown in Fig. 5. It essentially consists of matching the slope $p\ T\ R_{c1}/2N_GN_{c1}$ of the first summand to the slope of $f\ (i_{c1}N_{c1}+d)$, which is the core characteristic (Fig. 5(A)). Fig. 5(B) permits us to establish the values of d which are required, as well as the range over which variations of

$$E_1 = \frac{R_{c1}}{N_{c1}}(i_{c1}N_{c1})$$

can be regulated. N_G and p are found from the value of

$$E_2{}'\frac{T}{2N_G}$$

and of

$$\frac{T}{2N_G}\frac{R_{c1}}{N_{c1}}p$$

selected from the graphs, once R_{c1}/N_{c1} has been determined.

CASE 2

Since

$$E_1\frac{T}{2} = \text{const} = J_1 \tag{10}$$

equation 4 is rewritten as

$$E_2{}' = E_1\left(p - \frac{N_Gf\Sigma(i_cN_c)}{J_1}\right) \tag{11}$$

Rearranging equation 11 and substituting $R_{c1}/N_{c1}\ i_{c1c1}$ for E_1:

$$E_2{}'\frac{J_1}{N_G}\frac{N_{c1}}{R_{c1}} = p\ \frac{J_1}{N_G}\ i_{c1}N_{c1} - i_{c1}N_{c1}\times f(i_{c1}N_{c1}+d) \tag{12}$$

As with case 1, the optimum match of the slope $p\ J_1/N_G$ of the first summand to the slopes of the family of curves $i_{c1}N_{1c}f(i_{c1}N_{c1}+d)$ has to be found; see Figs. 6(A) and 6(B). In contrast to case 1, the curves for various values of d are not strictly parallel, which prohibits an equally good match for all values of d. As apparent from the graph of Fig. 6(A), a more uniform slope is found for negative values of d. Also, the slope $E_2{}'/d$ is steeper in this region.

For a perfect control characteristic, the function $f(i_{c1}N_{c1}+d)$ in equation 12 would have to be of the general form $f(x) = a - b/x$, which is a hyperbola with an asymptote at $f(x) = a$ corresponding to pJ_1/N_G, and an intersection with the x-axis at b/a, corresponding to

$$E_2{}'\frac{N_{c1}}{R_{c1}}\frac{1}{p}$$

in equation 12. Round HyMu 80-type core material approaches this desired characteristic better than others. The curves of Figs. 5 and 6 are based on data taken from 0.500- by 0.750- by 0.125-inch size 1-mil tape cores of this type of material.

According to equation 1, E_2 will drop if I_2 is increased, unless $E_2{}'$ is increased in the same way. The drop may be expressed as the difference ΔE_2 from a nominal voltage E_{20}.

$$E_2 = E_{20} - \Delta E_2 \tag{13}$$

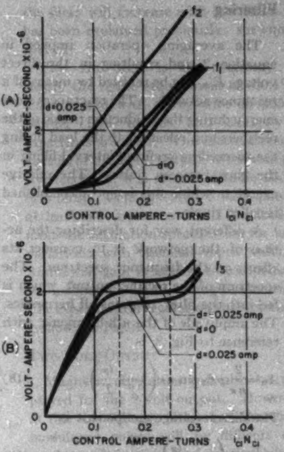


Fig. 6. Graphic representation of equation 12

$$f_1 = i_{c1}N_{c1}\ f(i_{c1}N_{c1}+d)$$

$$f_2 = p\frac{J_1}{N_G} i_{c1}N_{c1}$$

$$f_3 = E_2{}'\frac{J_1}{N_G}\frac{N_{c1}}{R_{c1}}$$

Applying this and equation 8 to equation 6:

$$d = \frac{E_{20}}{R_{c2}}N_{c2} - \frac{\Delta E_2}{R_{c2}}N_{c2} - \frac{E_R}{R_R}N_R \tag{14}$$

From equation 14 a suitable value of the reference ampere-turns, $(E_R/R_R)N_R$, may be selected, once a desirable value of d for nominal output voltage and current ($\Delta E_2 = 0$) has been determined from the graphs. For case 2, the voltage for $d=0$ and $I_2 = 0$ is best selected as the nominal output voltage because the curves of Fig. 6(B) show better regulation for negative values of d. Consequently, from equation 14

$$\frac{E_{20}}{R_{c2}}N_{c2} = \frac{E_R}{R_R}N_R \tag{15}$$

for the condition $\Delta E_2 = 0$ at $I_2 = 0$ and $d = 0$. Then $\Delta E \rightarrow 0$ when $I_2 \rightarrow 0$. Therefore, from equations 14 and 15

$$d = -\frac{\Delta E_2}{R_{c2}}N_{c2} \tag{16}$$

The required loop gain is determined from equation 1

$$\frac{\Delta E_2{}'}{\Delta E_2} = 1 + \frac{R_4\Delta I_2}{\Delta E_2} \tag{17}$$

where $\Delta I_2/\Delta E_2$ represents the output regulation requirement.

Filtering

The averaging operation implied in equation 2 and resulting in the direct voltage E'_2, can be realized by means of a reactance network. The reactances store energy during the conduction period of the rectifiers and release it to the load during the succeeding prefiring interval filling in the gaps between pulses. The relative extent to which this is not accomplished is called ripple.

A different way for describing the action of the network is to consider its effects on the frequency spectrum. The spectrum of the rectifier output, which is fed into the filter, contains all harmonics. The amplitude of the nth harmonic, with reference to Fig. 2, is

$$A_n = \frac{2}{n\pi} pE_1 \sin n\pi\left(1-2\frac{t_F}{T}\right) \qquad (18)$$

The direct-voltage component is

$$A_o = pE_1\left(1-2\frac{t_F}{T}\right) \qquad (19)$$

A_o is identical with E'_2 of equation 2. Consequently, the averaging network can be conceived as a low pass filter which suppresses all a-c components. The spectrum approach permits the design of the network attenuation characteristic to meet a specific steady-state ripple requirement.

Since the filter network is part of a closed-loop negative feedback system, the design also has to take into account the stability criterion. In its general form, it requires the complex loop gain to be less than unity at any frequency where its phase is π or odd integer multiples of π. In the combination magnetic-amplifier–low-pass filter, it suffices to make absolute loop gain less than unity before the phase shift becomes π, since the transfer characteristic of a magnetic amplifier resembles that of a low pass filter with monotonous high frequency roll-off.[6]

The simplest possible filter, for operation from a voltage source, consists of a large inductance in series with the load. The voltage source has to present a short circuit during the off period in order to maintain current flow, which enables the inductance to discharge its stored energy into the load. This essential requirement can be implemented by a diode across the output of the magnetic amplifier.

The performance of this simple filter may be improved if a capacitor is connected across the output. Its steady-state voltage transfer is now

$$\frac{E_1}{E_o} = 1 + d_R d_r - \left(\frac{f}{f_o}\right)^2 + j\frac{f}{f_o}(d_R + d_r) \qquad (20)$$

where

$$2\pi f_o = \frac{1}{\sqrt{LC}}$$

$$d_R = \frac{1}{R_L}\sqrt{\frac{L}{C}}$$

$$d_r = \frac{r}{\sqrt{\frac{L}{C}}}$$

The phase shift of this filter can never exceed −180 degrees, a desirable property with respect to the negative feedback stability. However, because of the rather gradual attenuation characteristic, the cutoff frequency, f_o, will have to be very low for adequate ripple reduction. This will slow down the regulation response, which can never be faster than $1/f_o$, and may result in large impractical values of inductance and capacitance.

A different approach[6] utilizes the magnetic amplifier to reduce the loop gain to less than unity before the phase has reached −180 degrees. The magnetic amplifier cutoff frequency is less than that of the filter and can easily be controlled through the effective value of R_e making the filter design less critical. It only has to consider the ripple requirement as long as its phase shift is kept low before the magnetic amplifier reaches unity gain. According to reference 6, the complex voltage gain of a full-wave magnetic amplifier for $fT<1/4(1/T=$ power frequency) is

$$G_v = \frac{G_{vo}}{(1+j\pi fT)\left(1+\frac{R_d}{R_c}\right)e^{j\pi fT/2}} \qquad (21)$$

G_{vo} is the voltage gain at zero frequency; R_e is the effective control resistance; and R_d is the trans-resistance of the core as defined in reference 6. The smaller the value of R_e, the lower will be the cutoff frequency of the magnetic amplifier. Because of the diode across the output, there is no interaction between the magnetic amplifier and the filter, and the loop gain is a product of both characteristics. The self-saturating magnetic amplifier will retard its firing angle to maintain a constant volt-second input to the filter if the diode ceases to conduct during the prefiring interval, because the inductor current becomes negative.

This condition will occur if the load current is:

$$I_2 < \frac{t_F E_2}{2L} \qquad (22)$$

A Practical 24-Volt to 9-Volt–2-Volt Converter

The design of this unit closely follows the concept analyzed in the preceding discussion. However, some of the specific engineering problems to be considered are not necessarily typical of this system but rather a result of the low output voltages. The converter is intended to supply base bias and collector voltage to high-speed germanium transistor switching circuits. The input and output specifications and the environmental conditions are listed in Table I.

Table I. Specifications for Converter

Input Voltage		Outputs			
Minimum	Maximum	Voltage	Regulation, %	Ripple, %	Maximum Current, Amperes
20	28	9	±4	<0.1	15
		2	±5	<0.1	2

Ambient temperature range: −5 to + 65 degrees centigrade.

Efficiency: No specific requirements. However, in the interest of simple thermal control of the unit itself, high efficiency is desirable.

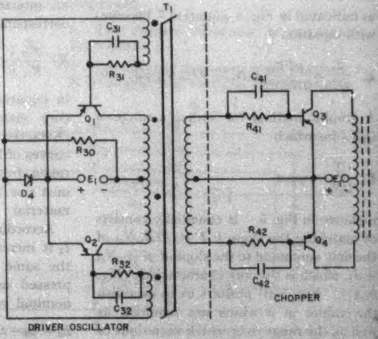


Fig. 7. Square-wave inverter circuit chopper

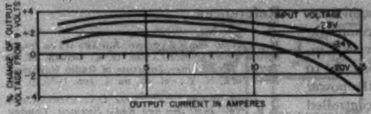


Fig. 8. Regulation of 9-volt output chopper

The Chopper Circuit

The principles of operation and design of the transistor-saturable-transformer square-wave oscillator have been covered extensively in literature. An important feature of this circuit is that the commutation of direct current into a square wave is controlled by a square-loop magnetic core resulting in constant and equal volt-second content of each half-wave. Since the transistors operate in the switching mode, high efficiency is obtained. The circuit shown in Fig. 7 provides a separate square-wave oscillator which drives the power transistors (Q_3, Q_4). In contrast to the conventional solution where the oscillator also handles the power, two extra transistors (of small power rating) and a small square-loop core have to be expended. However, this is well compensated for by the numerous advantages gained by the decoupling of the commutation control from the load. The most important advantages are:

1. Easy starting under full load.
2. Undisturbed commutation by the switching action of the regulator.
2. Very fast switching of the power transistors resulting in very low switching loss. For example, *2N1147A* transistors can switch 24 volts in 1 microsecond with an average current of 10 amperes. This is possible due to the better frequency characteristics of low-power transistors and the small thin-tape saturable core used in the driver.
4. The power-handling transformer is not operated into saturation. Consequently, it does not require square-loop core material. Core loss can be kept small and high saturation current transients do not occur.

The selection of the commutation frequency is a compromise between two conflicting aspects:

1. Low frequency means lower switching loss in the semiconductors and magnetic cores, but requires large magnetic cores and filter capacitors.
2. High frequency results in smaller physical size of magnetic core components and filter capacitors but entails greater switching loss.

In this case, the choice was 2,000 cycles per second for an input of 24 volts. With this frequency, core loss of the power transformer was found to be less than 3 watts. A 2-mil 50-50 nickel-iron tape core was used in the transformer. Since only a few turns of very heavy wire, such as six strands of no. *28* American Wire Gage for the 9-volt output, were required, this core was found to be the most economical. As deduced from the afore-mentioned switching time, the switching loss in the transistors amounts to less than 1 watt at maximum load current.

Since the driver oscillator consumes little power, it is possible to stabilize its frequency by means of a Zener diode circuit, which keeps the input voltage to the driver constant, such as previously described in case 1. However, adequate regulation also can be obtained if the frequency is allowed to vary with battery voltage as in case 2.

Nine-Volt Output

Externally, the circuit of Fig. 3 behaves exactly like a center-tapped full-wave magnetic amplifier. However, the gating of the output power is performed by a pair of silicon-controlled rectifiers, while the saturable reactors only have to trigger the rectifiers. Because of the very low trigger power requirements of these devices the magnetic cores are small and consume negligible power.[7] Consequently, power gain can be traded for faster response. The forward voltage drop of silicon-controlled rectifiers is practically the same as that of regular silicon rectifiers. Compared with a conventional magnetic amplifier the efficiency is improved by eliminating the I^2R losses of a gate winding and the losses of a large core. In the configuration shown, the trigger voltage is maintained only until the device has reached the conduction state. This reduces power loss even further and prevents reverse recovery effects in the diodes D_1 and D_2. However, if higher voltages have to be handled, the gate winding, N_G, can be connected to a lower voltage tap as indicated by a dotted line in Fig. 3.

Among other possible connections of the magnetic core reactors (for example, one across each half of the transformer, with the cathodes of the controlled rectifiers on the transformer side) only the one shown retains the essential characteristic for operation into a choke input filter with free-wheeling diode, D_3. It will adjust the firing angle if the diode leaves conduction too soon, which would maintain constant volt-second input to the filter. Otherwise, the output voltage of the filter will increase with low loads as has been explained previously; see equation 22.

Two *L-C* sections are necessary to satisfy the ripple requirement of <0.1%. Coils wound with heavy wire on 125 μ powder iron toroids and electrolytic capacitors are used. The internal resistance R_i of the 9-volt output, without negative feedback, was measured to be about 0.15 ohm. Consequently, a loop gain of at least 5.2 (equation 17) is required to meet the regulation specification for load variations with nominal input voltage. Additional gain is needed to compensate for incomplete input voltage regulation. The experimental regulation characteristic is shown in Fig. 8.

Two-Volt Output

Magnetic regulation can also be employed for the 2-volt output. However, at this low voltage, filtering of the width modulated pulses will be difficult and would result in the use of relatively large filter capacitors. Since the maximum required current is only 2 amperes, regulation by dissipation is feasible using the transistor circuit of Fig. 9. Its regulation characteristic (neglecting the small current passing through Q_5) is

$$E_2 = \frac{R_m}{R_n} E_R + \frac{R_m}{R_{51}h_5}\left(E_2' - e_6 - e_7 - I_2\left(R_{e7} + \frac{R_{51}}{h_6 h_7} + \frac{R_{e6}}{h_7}\right) + I_2\frac{R'}{R_n}\right) \quad (23)$$

where e_6 and e_7 denote the emitter to base threshold voltage (emitter current approaching zero) of the respective transistors; h_5, h_6, and h_7, signify the common emitter current gain; and R_{e6} and R_{e7} denote the emitter resistance.

The threshold voltage, e_6, e_7, is about 0.1 to 0.15 volt in germanium transistors and varies with temperature. It is normally neglected, but becomes important in this low-voltage circuit, unless E'_2 is made large which would result in large power dissipation. Consequently, R_m has to be small compared with $R_{51}h_5$. R' provides positive feedback reducing

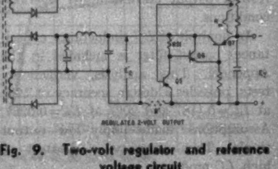


Fig. 9. Two-volt regulator and reference voltage circuit

the output current term in equation 23.[8] However, the practical circuit executed meets the regulation requirement of ±5% without resorting to positive feedback. The reference voltage supply, also shown in Fig. 9, serves both output circuits.

Conclusions

The design method presented has been successfully applied to a regulated 24-volt to 9-volt–2-volt converter. It permits optimization of the number of turns required by the saturable reactor of the 9-volt regulator circuit, which can be described as a full-wave magnetic amplifier with d-c output. Essentially, the method consists of a normalized graphic representation of the voltage transfer function of the entire system, which is based on the experimentally determined control characteristic of the saturable core from which the optimum conditions are readily deduced. The graphic approach is especially useful if the commutation frequency of the square-wave generator is not constant but proportional to the input voltage. Gating of the power is performed by a pair of silicon-controlled rectifiers. The saturable reactors are only used to provide the trigger timing for the controlled rectifiers. While the essential characteristics of the basic magnetic amplifier are retained, efficiency is greatly improved. Relatively little power is required from the 2-volt output. Therefore, a transistor regulator represents the most practical solution.

Because semiconductor and magnetic core devices are employed as the main elements, the described regulated d-c to d-c converter power supply can be expected to perform with a high degree of reliability. This is especially true if the thermal design, which was not covered in this paper, is carefully conducted.

References

1. A New Self-Excited Square-wave Transistor Power Oscillator, G. C. Uchrin, W. O. Taylor. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 43, no. 1, Jan. 1955, p. 99.

2. High Power Transistor D.C. Converters, T. R. Pye. *Electronic and Radio Engineer*, London, England, Mar. 1959, pp. 96–105.

3. Transistor Inverters and Rectifier-Filter Units, F. Butler. *Electronic Enineering*, London, England, July 1959, pp. 412–18.

4. The Controlled Rectifier—Key to the Continuing Control Renaissance, J. D. Harnden. "Proceedings, Special Technical Conference on Nonlinear Magnetics and Magnetic Amplifiers," *AIEE Special Publication*, 1958, pp. 414–24.

5. A Magnetic Amplifier-Silicon Transistor Power Supply for Missile Application, B. Mokrytzki, R. A. Stuart. *Ibid.*, *AIEE Special Publication T116*, 1959, pp. 179–95.

6. Time Response Constant and Transfer Function of Magnetic Amplifiers, E. W. Manteuffel. *Ibid.*, pp. 47–71.

7. Controlled Rectifier Manual (book). General Electric Company, Schenectady, N. Y., 1960.

8. Design of Transistor Regulated Power Supplies, R. D. Middlebrook. *Proceedings*, Institute of Radio Engineers, vol. 45, no. 11, Nov. 1957, pp. 1502–09.

A 30-Volt 10-Ampere Regulated Power Supply Using the Variable Output Transistor Oscillator with Current Transformer

S. P. JACKSON
MEMBER AIEE

A PREVIOUS unpublished paper by the author derived the design equations for the transistor saturable current transformer amplifier. In that discussion, sample calculations were made and an amplifier designed. Measured values were compared with those calculated. It is a short step to use the amplifier previously designed in a direct output power supply to provide a constant 30 volts up to 10 amperes. A current limiting circuit was added to the circuit to restrict output current under overload conditions.

An analysis of *LC* (inductance capacitance) output filter is included in the Appendix. This analysis, verified by tests, provides a design accuracy of 12% at $LC=0.1187$ and 20% at $LC=0.0267$. Assumptions made imply low output ripple. Thus accuracy is greatest at high *LC* products.

Circuit Operation

The complete circuit is shown in Fig. 1. Square output voltage pulses of constant amplitude and width (with constant applied voltage and constant load resistance) and varying off-time are provided by the *DT100* and saturating transformer. Off-time is varied by changing the reset voltage across N_r.

At the beginning of the cycle the supply voltage appears across N_p, the load (through an *LC* filter) and the *DT100* emitter-collector. The voltage appearing across N_p is equal to the primary current multiplied by the emitter-base and base resistance of the *DT100* reflected to the primary. This voltage takes the transformer to saturation. At this point the windings N_p and N_r are effectively decoupled. The reset voltage supplied by E_r through the series transistor *1724* now functions to reset the core to negative saturation. At the same time this voltage back-biases the base of the *DT100*, opening the load circuit by means of the transistor switch. The time it takes to reach negative saturation varies with the amount of reset voltage applied across N_r. This reset voltage is changed by means of a feedback loop in a manner to maintain load voltage constant.

The regulator portion of the circuits consists of a current source (130-volt d-c supply and 4-megohm resistor), reference and detector (*SV3292*), and amplifier regulator (*366* and *1724*). Constant current flows to the junction of the *366* base and *SV3292* collector. If the load voltage is high the base-emitter voltage of the *SV3292* is high, increasing the amount of the current taken by the *SV3292*. This in turn reduces the current to the base of the *366*, causing more voltage to appear across emitter to collector of the *1724*. Thus less reset voltage is applied across N_r and a longer off-time occurs, reducing the output load voltage. If the load volt-

---

Paper 61-717, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted November 1, 1960; made available for printing April 19, 1961.

S. P. Jackson is with the Continental Electronics Corporation, Columbus, Ohio.

The author gratefully acknowledges the help of Dr. W. Herbert Bixby of the Power Equipment Company and the support of the Power Supply Group of the International Business Machines Corporation.

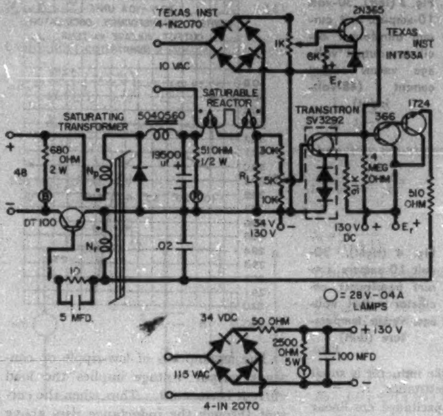


Fig. 1. Schematic wiring diagram of 30-volt 10-ampere unit

age tends to be low, a reverse process occurs to reduce off-time and increase average load current.

Current limiting is provided by the series-connected saturable reactor with a relatively high control impedance (d-c measuring reactor). Current flows from the 10-volt a-c source proportional to direct load current. Thus the voltage on the base of the *2N365* is proportional to load current. At some preset value of load current, the base voltage exceeds the emitter voltage of the *2N365* fixed by the regulator diode *1N753A*. At this point, collector current flows to the *2N365* reducing the base current to the *366* and thereby increasing off-time and reducing average load voltage.

Circuit Results

If, instead of the regulating system, a simple manually controlled direct voltage is applied to the reset winding, d-c to d-c efficiency of slightly over 90% is obtained at rated load conditions (30 volts, 10 amperes). The addition of the feedback system reduces this efficiency to 82%. Thus the efficiency of the power amplifier using the saturating core and switching transistor is about equal to that of the comparators, amplifiers and the d-c transformer. Note that the efficiency defined is d-c to d-c efficiency. Further over-all reduction would occur if conversion efficiency is desired.

It was the aim of this development to determine the minimum volume of a 30-volt 10-ampere constant voltage supply with current limit. Accordingly the filter capacitor was selected on the basis of available volume rather than the degree of filtering or the dynamic response. The selection of the 19,500-microfarad (μf) filter capacitor yielded a peak-to-peak ripple of 22 millivolts at rated load. Volume of the filter choke was reduced by using a "swinging" type. It is important to reduce volume. The fact that off-time increases at light loads makes this approach desirable.

A means of determining the size of inductance and capacitance for the filter to yield a specified ripple is included in the Appendix. These curves, Fig. 1, may also be used to determine the inductance as a function of current for a given filter capacitor if the main supply voltage is essentially constant. This is easily done by determining η, the ratio of conducting or on-time to nonconducting or off-time as a function of load current.

With the filter noted, the dynamic response to 10% line and load changes in either direction, singly or jointly, showed no detectable overshoot. Response time of the power amplifier may be varied by operating frequency. This is easily adjusted to provide closed loop stability. No stability problems were noted in this design.

Fig. 2 indicates the degree of output voltage regulation with changing input voltage. All supplies were varied simultaneously, however, only the main input supply voltage is used in the plot. As might be expected, the load range of the unit is sensitive to line changes. The fact that a saturating current transformer is used to switch the transistor (*DT 100*) in the power amplifier leads one to expect such results for supply variations. A two to one load change is about the maximum allowable variation for ±10% supply voltage change. This points to the need for better regulation of the input supply.

Load voltage regulation versus load current is displayed in Fig. 3 for two different current cut-off conditions. Close regulation of output voltage is maintained from about 36% to 100% of rated load current with constant supply voltage. While greater range was not necessary in this design, it is believed that a more stable biasing arrangement in the preamplifier would yield a substantially larger range.

No attempt was made to reduce drift of output voltage other than the use of the matched combination of reference and first stage amplifier. Representative drift is shown in Fig. 4. The unit as built is satisfactory to a maximum temperature of about 43 C (degrees centigrade). Stud temperature rise of the swiching transistor at this temperature was 26 C. The volume of the packaged unit was 250 cubic inches or about 0.833 cubic inch/watt.

Conclusions

The circuit discussed takes advantage of the ability of transistors to handle

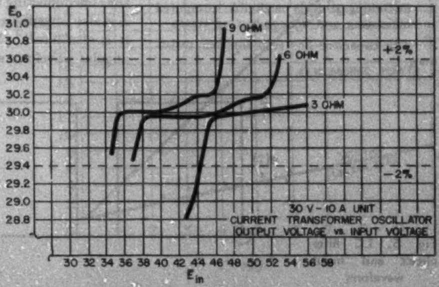


Fig. 2. 30-volt 10-ampere unit current transformer oscillator output voltage versus input voltage

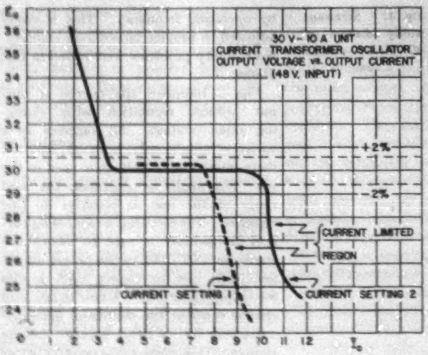


Fig. 3 (left). 30-volt 10-ampere unit current transformer oscillator output voltage versus output current (48-volt input)

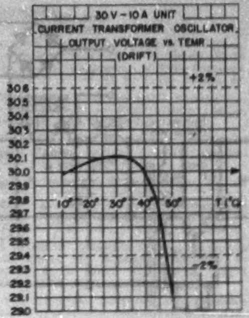


Fig. 4 (right). 30-volt 10-ampere current transformer oscillator output voltage versus temperature (drift)

large amounts of power when used as switching devices. It further provides short response times by the use of short conducting periods. Since little power is converted to heat by the switching transistor, low stud-temperature rise results in greater safety to the transistor. Also, high d-c to d-c efficiency is possible.

It is not intended that this circuit be represented as a finished design. In this paper an attempt is made to show some of its advantages by means of an example.

Appendix

The transistor oscillator circuits provide an output voltage which is essentially a square pulse. Both the conducting and nonconducting times may vary.

In order to provide a low-ripple direct-voltage output, a filter must be provided. A typical circuit is shown in Fig. 5.

Calculation of the value of inductance and capacitance necessary to achieve a desired ratio of peak-to-peak ripple voltage to direct output voltage is the problem.

Assumptions

1. Supply source has zero impedance.
2. Switching time is negligible compared to conducting and nonconducting times.
3. The capacitor has zero internal resistance.
4. The resistance of the inductor is small compared to the load resistance.
5. Inductance and capacitance are linear for a given load condition (over the peak-to-peak ripple current range).
6. Peak-to-peak ripple voltage is small compared to voltage across inductance.

Analysis

Consider the choke input filter circuit shown in Fig. 6 where $e(t)$ is of the form illustrated in Fig. 7. E_s is greater than E_o by an amount sufficient to provide energy storage in the inductance and capacitance which is given up to the load during the nonconducting time. Assuming the ripple voltage is small compared to the difference between E_s and E_o during the conducting time, the voltage across the inductance is a rectangular wave (Fig. 8). Thus during the conducting interval:

$$L\frac{di_{Lc}}{dt}=E_s-E_o$$

$$\frac{di_{Lc}}{dt}=\frac{E_s-E_o}{L} \qquad (1)$$

Since $\bar{E}_s-\bar{E}_o/L$ is a constant, the current through the inductance is linear. During the nonconducting time, the voltage across L is also constant and the current through it is also linear. We thus have the triangular waveform shown in Fig. 9.

The assumption of low-ripple or constant output voltage implies the load current is constant. Thus when the current through the inductance rises above the average value, the difference between its value and load current flows to the capacitor. When it falls below the average value, the capacitor must supply a portion of load current. Over the interval when the current through the inductance is greater than the average load current, shaded area, the peak current equals $di_{Lc}/dt\, t_c/2$ and the added capacitor charge is:

$$\Delta q=\frac{1}{2}\left(\frac{di_{Lc}}{dt}\frac{t_c}{2}\right)\left(\frac{t_c}{2}+\frac{t_n}{2}\right) \qquad (2)$$

On combining equations 1 and 2:

$$\Delta q=\frac{t_c}{8}(t_c+t_n)\left(\frac{E_s-E_o}{L}\right) \qquad (3)$$

The capacitor voltage increase above the average value due to Δq is equal to half the peak-to-peak ripple voltage.

$$\frac{\Delta E_o}{2}=\frac{\Delta q}{C} \qquad (4)$$

Combining equations 3 and 4:

$$\Delta E_o=\frac{2}{C}\left[\frac{t_c}{8}(t_c+t_n)\frac{E_s-E_o}{L}\right]$$

$$=\left(\frac{E_s-E_o}{4LC}\right)t_c(t_c+t_n) \qquad (5)$$

But

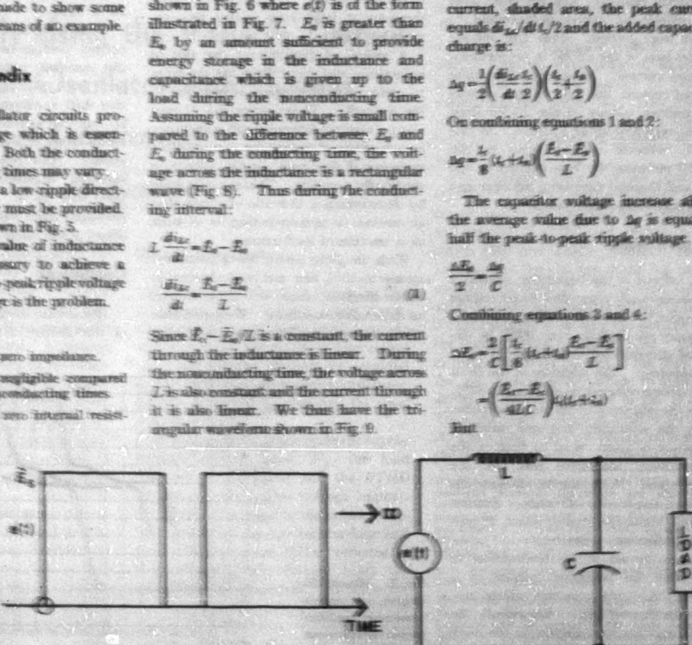


Fig. 5. LC filter circuit and input waveform

$$\bar{E}_o = \hat{E}_s \frac{t_c}{t_c + t_n} \qquad (6)$$

Combining equations 5 and 6

$$\Delta E_o = \frac{\bar{E}_o}{4LC}\left[\left(\frac{t_c+t_n}{t_c}\right)-1\right](t_c+t_n)t_c$$

$$= \frac{\bar{E}_o}{4LC}\left[\frac{(t_c+t_n-t_n)(t_c+t_n)t_c}{t_c}\right]$$

$$\Delta E_o = \frac{\bar{E}_o}{4LC} t_n(t_c+t_n) \qquad (7)$$

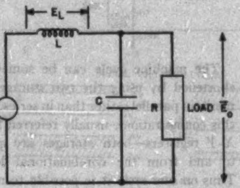


Fig. 6. Nomenclature for choke input filter

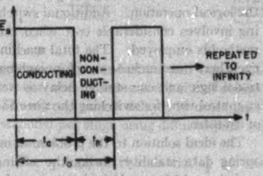


Fig. 7. Nomenclature for input waveform

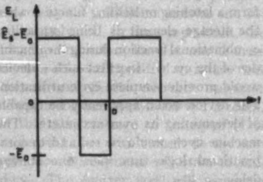


Fig. 8. Inductance voltage waveform

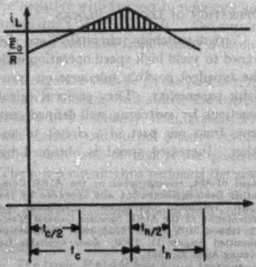


Fig. 9. Inductance current waveform

If

$$\eta = \frac{t_c}{t_n} \qquad (8)$$

and

$$f = \frac{1}{t_o} = \frac{1}{t_c - t_n} \qquad (9)$$

then from equations 8 and 9

$$t_c = \eta t_n = \frac{1 - f t_n}{f}$$

$$t_n = \frac{1}{f(\eta+1)} \qquad (10)$$

Combining equations 9 and 10

$$\Delta E_o = \frac{\bar{E}_o}{4LC}\left[\frac{1}{f^2(\eta+1)}\right] \qquad (11)$$

and

$$\frac{E_o}{E_o} = \frac{1}{LC}\left[\frac{1}{4f^2(\eta+1)}\right] = \frac{\Gamma}{LC} \qquad (12)$$

CRITICAL INDUCTANCE

The critical inductance is that value of inductance which insures that current flows in the inductance throughout the cycle. Thus the critical inductance is defined by letting $i_L = 0$ at the end of the nonconducting time.

Therefore

$$\frac{\bar{E}_o}{R} + \frac{di_{Ln}}{dt}\frac{t_n}{2} = 0 \qquad (13)$$

Since during the nonconducting time $\hat{E}_s = 0$ the voltage across L is $\bar{E}_o$ and

$$\frac{di_{Ln}}{dt} = \frac{\bar{E}_o}{L_o} \qquad (14)$$

Combining equations 13 and 14

$$\frac{\bar{E}_o}{R} = \frac{\bar{E}_o}{L_c}\frac{t_n}{2}$$

and

$$\frac{L_o}{R} = \frac{t_n}{2} \qquad (15)$$

or combining equations 10 and 15,

$$\frac{L_c}{R} = \frac{1}{2f(\eta+1)} = 2f\Gamma \qquad (16)$$

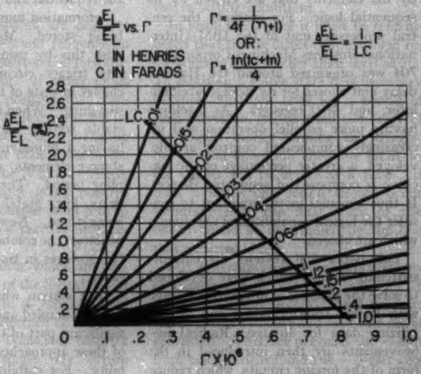


Fig. 10. Peak-to-peak ripple (expressed in per cent of direct output voltage) versus Γ with LC product as a parameter

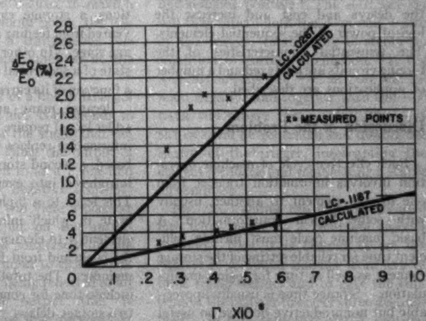


Fig. 11. Experimental verification of filter analysis

Results

The output peak-to-peak voltage may be predicted by

$$\frac{\Delta E_o}{E_o}=\frac{t_n(t_c+t_n)}{4LC}=\frac{\Gamma}{LC}=\left[\frac{1}{LC}\frac{1}{4f^2(\eta+1)}\right]$$

A curve of $\Delta E_o/E_o$ versus Γ for various values of LC is included (Fig. 10).

Critical inductance is defined by

$$\frac{L_c}{R}=\frac{t_n}{2}=2f\Gamma=\frac{1}{2f(\eta+1)}$$

The correlation between calculated and measured values is shown in Fig. 11.

Nomenclature

C = filter capacitance
$e(t)$ = applied instantaneous voltage
$\hat{E}_s$ = peak of square wave applied voltage
E_L = voltage across inductance
$\bar{E}_o$ = average value of load voltage
ΔE_o = peak-to-peak ripple voltage across load
f = frequency
i_{Lc} = instantaneous current through inductance during conducting time
i_{Ln} = instantaneous current through inductance during nonconducting time
L = filter inductance
L_c = critical filter inductance
Δ_q = change in capacitor charge
R = load resistance
t = time
t_n = nonconducting time
t_c = conducting time
t_0 = period
η = ratio of conducting to nonconducting time
$\Gamma = t_n(t_c+t_n)/4$

Forcing Circuitry: Sequential Building Blocks for Logical Design

R. M. MEADE
NONMEMBER AIEE

THE PERFORMANCE of an entire computer system is greatly influenced by the elements that are chosen for its sequential logic. For example, the central processing unit of the IBM (International Business Machines Corporation) 704 was organized around the Haven's delay unit chosen for its storage, accumulator, and *M-Q* registers. As machines having more parallel operation and increasingly complicated controls are developed, a greater percentage of the machine is composed of sequential devices. It thus becomes more important to consider the logical capabilities of the circuits which are chosen.

This paper first reviews the current-switching circuits which were developed to solve the problem of integrating storage with logical transformation and of stabilizing data for storage. Radical improvements are then introduced in the form of the forcing circuits, which reduce the delays and cost and increase the logical power of the sequential elements. The sequential characteristics of the forcing circuits are analyzed and a number of applications are described.

Discussion of the Problem

Each effective cycle of machine operation involves information transfer from one storage element to another, usually with some logical transformation. A basic machine cycle must include sufficient time for reliable setting of the storage element as well as time for logical manipulation. Storage time is usually appreciable but nonproductive because no useful transformation is taking place. Speed of operation can be increased by integrating the sequential and combinational logic.

Information must be stable while it is being stored. Many systems accomplish this by using capacitive coupling into trigger circuits, so that only the change in state of the inputs can produce switching. When direct coupling is used, stable setting may be guaranteed by having a minimum path delay, between storage elements, greater than the sampling period. The logic then acts as a delay line, stabilizing the information being stored. New information can never penetrate the combinational logic between two storages in time to be gated erroneously.

In a system where logical blocks are directly coupled and the storage time is a significant part of the basic cycle, neither of these approaches will insure data stability. For stability under these conditions, a storage element must be prevented from feeding itself as the next storage stage. In other words, the change in state of a sequential circuit must never be a function of its current state.

Because many applications (such as adder loops) require that newly generated information replace that on which it was based, a second storage element must be introduced into every information loop. This leads to a 2-phase or *A-B* machine cycle in which information flows from element *A* to element *B* in one portion of the cycle and from *B* to *A* during the remainder. The total machine cycle must include time for combinational logic plus two storage delays.

The machine cycle can be somewhat shortened by using the two storage elements in parallel rather than in series. In this configuration—usually referred to as *X-Y* registers—both storages are gated to and from the combinational logic. Thus on one cycle it is possible to make a transfer from register *X* through the logic to register *Y* and on the next cycle to reverse the transfer, again performing the logical operation. Additional switching involves considerable cost when this method is employed. The total machine cycle here must include time for combinational logic and one storage delay as well as control time for switching the direction of transfer.

The ideal solution to the problem of insuring data stability during the setting period would be to design a storage element that could be part of the combinational logic. This element should perform a latching or holding function when the storage element is being set and a combinational function during the remainder of the cycle. In effect such a device would provide complete cycle utilization. The device would apparently be capable of determining its own next state. The machine cycle would be reduced to combinational logic time plus one storage delay.

Current-Switching Circuits

OPERATION OF LOGICAL BLOCKS

Current-switching circuits[1-3] are designed to yield high-speed operation with the broadest possible tolerance on transistor parameters. They perform logical functions by switching well-defined currents from one part of a circuit to another. Increased speed is obtained by

---

Paper 61-814, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted August 11, 1960; made available for printing April 25, 1961.

R. M. MEADE is with International Business Machines Corporation, Poughkeepsie, N. Y.

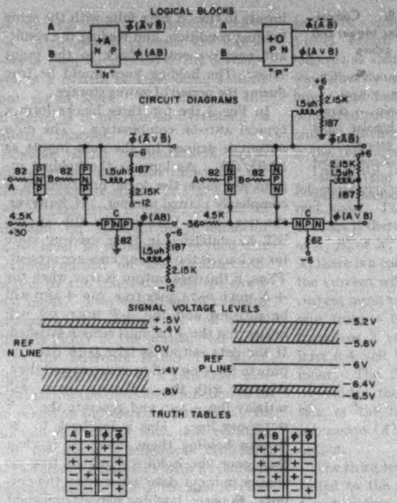


Fig. 1. Current-switching circuits

operating the transistors well below saturation and by using small signal swings. The average circuit delay is 20 nsec (nanoseconds) per stage using diffused-junction transistors.

The principal characteristics of the current-switching logical blocks are shown in Fig. 1. The common-emitter junction is fed by a 6.6-mil current source represented by the 4.5-kilohm resistor and 30-volt supply. If, in the p-n-p circuit, the two top transistors have *A* and *B* base inputs sufficiently positive to insure that they are cut off, the reference transistor *C* will be forward-biased and will conduct. If the input voltage to transistor *A* is driven negative with respect to the zero reference, transistor *A* will start to conduct. The emitter node will assume a voltage which is forward-diode-drop above the conducting base. It will, however, be negative with respect to reference and transistor *C* will be cut off. The current has been switched from output ϕ to output $\bar{\phi}$.

A negative voltage on the base of transistor *B* similarly will switch the current from ϕ to $\bar{\phi}$. Clearly, then, transistor *C* will conduct and output ϕ carry current only when both *A* and *B* are positive. The p-n-p circuit thus performs the positive AND function. The outputs ϕ and $\bar{\phi}$ are inverse functions, bipolar on a voltage basis and complementary on a current basis.

The basic *P* block has n-p-n transistors connected in a similar configuration. If both *A* and *B* transistors are negative with respect to the −6-volt reference on transistor *C*, then transistor *C* is forward-biased and conducting. If either input *A* or *B* is now driven positive, that transistor will conduct and the current will be switched from output ϕ to $\bar{\phi}$. During this switch the emitter node shifts in the same manner as for the *N* block. Since, in this case, either input going positive causes the switch in output from ϕ to $\bar{\phi}$, the *P* block performs a positive OR function.

The coupling networks which are connected to the outputs of the logical blocks convert the current to a voltage signal swinging about the reference of the complementary circuit. Transistor bases of succeeding logical stages are driven directly. When it is desired to drive a load located at a remote point, the collector current feeds a transmission line having a network and amplifier at the receiving end.

It is usual to speak of a line in terms of its reference level, which corresponds to the reference level for the stages that can be directly driven. Thus, the output of an *N* block is referred to as a *P* line and drives *P* blocks. A negative (current) input to an *N* block will yield a positive (current) *P*-line output on the out-of-phase ($\bar{\phi}$) output and conversely for the *P* block. A current input to either block may be thought of as being reflected as an out-of-phase current output that is translated in reference and reversed in polarity; both the *N* and the *P* blocks are current ORS. The operation of more involved circuits can be described more conveniently in terms of current logic, which yields identical equations for p-n-p and n-p-n circuits.

Current-switching logical circuits form a convenient base for the design of sequential circuits. Because the circuits have a d-c stability factor of unity and each stage generates a reset output level, any logically meaningful feedback connection can be made. Direct couplings and the switching of controlled currents reduce susceptibility to noise.

STORAGE USING CURRENT SWITCHING

The basic current-switching flip-flop consists of a *P* block and an *N* block cross-coupled to form a full loop as shown in Fig. 2. The feedback results in a bistable trigger circuit. The operation of the blocks which comprise the trigger is identical with that of the basic circuits discussed before. As illustrated in Fig. 2(A), inputs may be added to one or both sides of the trigger to provide for alternative setting conditions.

In using the trigger it must be remembered that this is direct-coupled logic responding to levels rather than level shifts. To establish a stable state it is necessary to apply an input level for a period of time somewhat greater than the delay of the logical blocks forming the trigger. For normal operation only one input to the trigger will be active at a time. If both a set-on and set-off should occur simultaneously, the final state of the circuit would be determined by the one that persisted longer.

Fig. 2(B) shows a practical arrangement of trigger and bipolar input gating. If the +*N* BIT and +*N* GATE are present, the +AND is satisfied and +*P* SET ON is generated. This line prevents the −AND from being satisfied and the −*N* SET OFF line remains positive. The +*P* SET ON will then satisfy both the +OR and +AND of the trigger, establishing a stable on condition of current circulation. If the +*N* BIT is false the +*P* SET ON will remain negative and together with the −*P* GATE will satisfy the −AND. The −*N* SET OFF will be generated and disable the +AND of the trigger, interrupting the current feedback and establishing a stable off condition.

Fig. 2(C) shows the trigger extended into what might be a complete register position. Here it is desired to set the register in a bipolar manner from any of three signal sources and to gate the contents of the register to one of three destinations. In this case two of the destina-

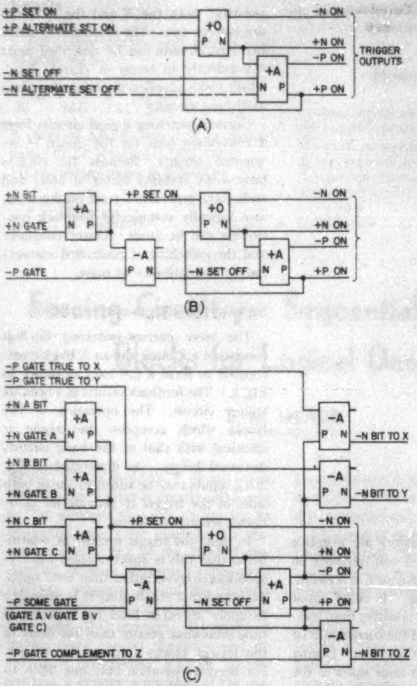

+P SET ON
+P ALTERNATE SET ON
−N SET OFF
−N ALTERNATE SET OFF
+O
P N
+A
N P
−N ON
+N ON
−P ON
+P ON
TRIGGER OUTPUTS
(A)
+N BIT
+N GATE
−P GATE
+P SET ON
−N SET OFF
+A
N P
−A
P N
+O
P N
+A
N P
−N ON
+N ON
−P ON
+P ON
(B)
−P GATE TRUE TO X
−P GATE TRUE TO Y
+N A BIT
+N GATE A
+N B BIT
+N GATE B
+N C BIT
+N GATE C
−P SOME GATE
(GATE A V GATE B V GATE C)
−P GATE COMPLEMENT TO Z
+A
N P
+P SET ON
−N SET OFF
+O
P N
−A
P N
−N BIT TO X
−N BIT TO Y
−N BIT TO Z
−N ON
+N ON
−P ON
+P ON
(C)

Fig. 2. Current-switching trigger and gating

A—Current-switching trigger
B—Single-directional bipolar input gating
C—Multiple input and true/complement output gating

tions require the true contents and one (probably an adder) requires the complement.

This arrangement introduces the concept of a wire OR. If the outputs of two or more blocks are mutually exclusive in a current sense, it is possible for them to share a coupling network. Thus, in this case, there are three +AND circuits with their in-phase outputs tied together. When a +AND is satisfied, the in-phase output will conduct; to be mutually exclusive in a current sense, one +AND circuit only must be satisfied. Since only one input to a register will be gated on a single cycle, the wire OR criterion is met. The ability to wire OR makes it possible to share a common −AND to set the register off. To utilize this feature, however, a gating signal that is an OR of all the individual gates must be generated.

DATA-LATCHING FUNCTION

As previously discussed, the data to be stored in the current-switching trigger must be stable during the setting period. Rather than use a second trigger, a method was developed for integrating a holding function with combinational logic. This current-switching data latch is shown in Fig. 3.

It was first observed that most combinational logic (e.g., adder outputs) ends in an AND–OR configuration. It was then realized that the OR could be held true during the gating period by feeding back its own output, ANDed with a holding condition. Similarly, the circuit could be held false by ANDing the fact that both inputs to the OR were false with the same holding condition, and feeding the resultant negative output back to the input ANDs. The holding lines would be true during the period of gating storage.

In Fig. 3 the top three blocks form a typical AND–OR combination. The data output is defined by the data inputs as $(AB\vee CD)$. As long as the holding inputs are false the logic functions in a completely normal manner. If, however, the HOLD DATA lines go true, the outputs will be stabilized in their existing state for as long as the holding lines are present. Thus, if the data output is true when the $+N$ HOLD DATA goes true, the +AND will be satisfied and the $+P$ HOLD ON will insure that the +OR must remain satisfied. If the data output is false both data inputs to the +OR will be negative and they, together with the $-P$ HOLD DATA, will satisfay the −AND and generate the $-N$ HOLD OFF line. This is fed back to the +ANDs holding them false. It is clear that once the holding lines go true no change in input data will modify the output. Between latching periods sufficient time must be allowed for new data to flow through the latch.

The $+P$ HOLD ON and $-N$ HOLD OFF lines can be directly used for switching a current-switching trigger. The dotted lines in Fig. 3 show outputs for this purpose. When this is done it is necessary to add the dotted input to the −AND circuit. This prevents a latched true condition followed by the data going false from generating both SET ON and SET OFF lines.

Forcing Circuitry

THREE-LEVEL LOGIC

The essential principle of current-switching logic is the steering or switching of a constant current into a path defined by the collector of the transistor having the most negative (p-n-p) or positive (n-p-n) base in the circuit. As described above, it is a 2-level system in which bases assume a value positive or

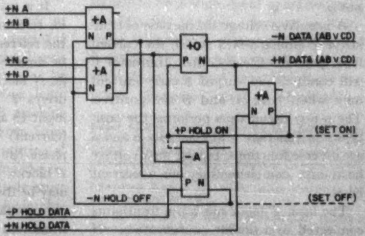

+N A
+N B
+N C
+N D
+A
N P
+O
P N
−N DATA (AB V CD)
+N DATA (AB V CD)
+A
N P
+P HOLD ON
(SET ON)
−A
P N
−N HOLD OFF
(SET OFF)
−P HOLD DATA
+N HOLD DATA

Fig. 3. Current-switching data latch

negative with respect to a specified reference. Attempts to increase the logical power of these circuits led to the development of a system of logic using a third input signal level.[3] The effect of the third level is to provide an overriding or inhibitor[4] function.

Possible circuit configurations, signal voltage levels, and truth tables for the 3-level blocks are shown in Fig. 4. The third voltage level is chosen so that its minimum "on" (plus for n-p-n, minus for p-n-p) level is sufficiently beyond the maximum normal "on" level to insure conduction by the transistor with the third-level input. Two variations of coupling network exist, yielding an "off" level which is either at the normal "off" level or at the reference level. The latter, which is referred to as a split-level network, allows the third-level function to be impressed on the reference transistor. Its function is identical with that resulting from the use of independent third-level and reference transistors with a common-collector output. Use of the split level reduces switching noise in the logical block.

The p-n-p circuit shows a third-level input and the n-p-n circuit shows a split-level input, but both versions are used for either circuit. An arrow head is used to designate a third-level input, an *S* to designate a split-level input. In drawing 3-level logical blocks the convention followed is that collector outputs are always directly opposite corresponding base inputs and common collectors are shown connected outside the block. The bottom output, having no corresponding input, is the reference transistor output.

The truth tables in Fig. 4 and the following explanation are in terms of current logic. The reference transistor will conduct if there is no current input to any of A, B, or X. If now, with no input to X, there is a current input to either A or B the current will switch from ϕ to $\bar{\phi}$. A current input to X will now switch the current from ϕ to X because the base of transistor X receives an input that is at least 0.4 volt beyond either the A or B input. The third-level input has thus overridden the normal current OR function of the block, yielding the logical expression $(A \vee B)\bar{X}$ for the out-of-phase output.

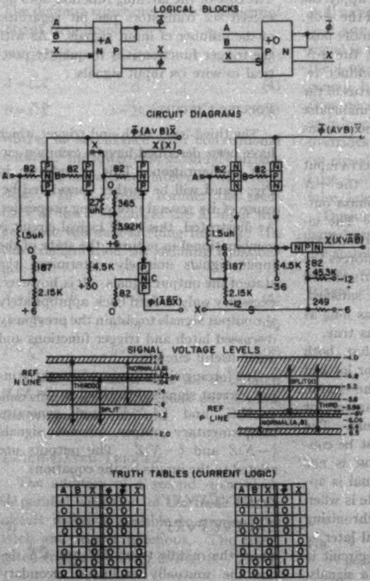
The third-level technique is particularly suited to the problem of multiple directional gating. A multiple AND or distribution function can be obtained by using a circuit having a single-bit input at the third level and a number of gating inputs at the normal level, each with its collector independently loaded. If the bit signal is defined as $+NX$ and the gating signals as $-NA_i$, the expressions for the gated bit are $O_1 = XA_1$, $O_2 = XA_2$, $\ldots$, $O_n = XA_n$, where no two gates can occur simultaneously. The gating signals must be mutually exclusive to prevent distribution of current through the independent collectors.

Application of Third Level to Triggers

The third-level concept was first applied to sequential circuits in developing a method of bipolar input gating to triggers. This logic is illustrated in Fig. 5. It is accomplished by a combination of the P and N blocks of Fig. 4.

The operation of the trigger will be described starting with an initial state in which the trigger is off, all outputs are false, and the $+N$ GATE is absent. Because the negative third-level GATE signal forces conduction by that transistor, neither $+P$ SET ON nor $+P$ SET OFF can be true. The trigger can be switched "on" by the $+N$ GATE signal going true with the $-N$ BIT present. The absence of the current into the third-level input of the $+$AND allows the transistor having the $-N$ BIT input to conduct. This conduction causes $+P$ SET ON to go true while $+P$ SET OFF remains false. The $+$OR responds to $+P$ SET ON in a normal manner yielding the $\pm N$ ON signals. The single-input $+$AND is satisfied by the $+N$ ON causing the $\pm P$ ON signals to become true. The $+P$ ON is fed back into the $+$OR of the trigger, holding that circuit in the on condition. The gating signal can now be removed, dropping the $+P$ SET ON but leaving the trigger in the on condition.

In a similar manner the trigger can be switched back to the off condition by reapplying the gating signal with the $-N$ BIT absent. In this case the $+N$ GATE and the absence of the $-N$ BIT satisfy the


LOGICAL BLOCKS
A
B
X
+A
N P
$\bar{\phi}$
X
ϕ
A
B
X
S
+O
P N
$\bar{\phi}$
X
CIRCUIT DIAGRAMS
$\bar{\phi}(A \vee B)\bar{X}$
X
X(X)
A
82
B
82
82
P N P
P N P
P N P
27 uh
O
365
3.92K
+6
1.5uh
O
187
2.15K
+6
4.5K
+30
P N P
C
82
O
$\phi(\bar{A}\bar{B}\bar{X})$
$\bar{\phi}(A \vee B)\bar{X}$
A
82
B
82
N P N
N P N
1.5uh
-6
187
2.15K
-12
4.5K
-36
N P N
$X(X \vee \bar{A}\bar{B})$
82
45.3K
-12
249
-6
X
S
SIGNAL VOLTAGE LEVELS
REF N LINE
NORMAL(A,B)
THIRD(X)
SPLIT
+.5
+.4
+.04
0V
-.04
-.4
-.8
-1.2
-2.0
REF P LINE
SPLIT(X)
THIRD
NORMAL(A,B)
-4.0
-4.8
-5.2
-5.6
-5.96
-6V
-6.06
-6.4
-6.5
TRUTH TABLES (CURRENT LOGIC)

A	B	X	ϕ	$\bar{\phi}$	X
1	1	1	0	0	1
0	1	1	0	0	1
1	0	1	0	0	1
0	0	1	0	0	1
1	1	0	0	1	0
0	1	0	0	1	0
1	0	0	0	1	0
0	0	0	1	0	0

A	B	X	$\bar{\phi}$	X
1	1	1	0	1
0	1	1	0	1
1	0	1	0	1
0	0	1	0	1
1	1	0	1	0
0	1	0	1	0
1	0	0	1	0
0	0	0	0	1

Fig. 4 (left). Three-level current-switching circuits

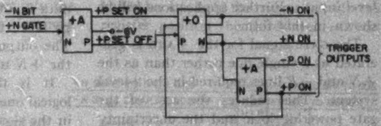

-N BIT
+N GATE
+A
N P
+P SET ON
-6V
+P SET OFF
+O
P N
-N ON
+N ON
+A
N P
-P ON
+P ON
TRIGGER OUTPUTS

Fig. 5 (right). Third-level gated trigger

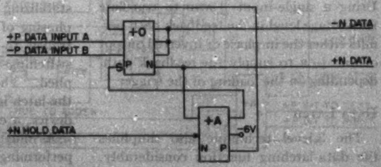

+P DATA INPUT A
-P DATA INPUT B
+O
S P N
-N DATA
+N DATA
+N HOLD DATA
+A
N P
-6V

Fig. 6 (right). Third-level data latch

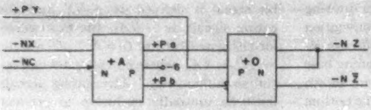


Fig. 7 (left). Combinational forcing circuit

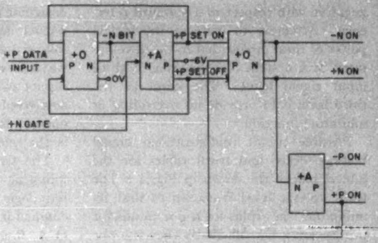


Fig. 8 (right). Latch-gated trigger

+AND, so that the reference transistor conducts yielding the $+P$ SET OFF. Since this signal is at the third level it overrides the $+P$ ON that had been latching the trigger "on" and switches the +OR "off." The $\pm N$ ON outputs become false causing the $\pm P$ ON outputs to change from true to false and to remove the signal which had been attempting to hold the trigger "on." If the gating signal is removed, the $+P$ SET OFF is removed and the circuit reverts to its off state.

Two important consequences of the third-level approach are apparent in a comparison of Figs. 2(B) and 5. The first is the elimination of the −AND gating block required in the 2-level approach. The second is the reduction of signal delay through gate and trigger from three to two logical stages for the fast output. The 2-level approach requires three stages of delay to obtain a symmetrical delay path for the turn-off and turn-on signals. If the outputs of the +OR were used, the turn-on delay would be two stages but that for turn-off would be four.

The data bit feeding the third-level trigger was defined as a $-N$ line. The only consequence of having a $+N$ data line would be to redefine the in-phase output and the +AND as $+P$ SET ON and the inverted output as $+P$ SET OFF and to require their reversal as inputs to the +OR. As in the 2-level version, additional gates may be wire ORed into the setting lines. Because the collector of the transistor in the +OR having the third-level input is tied to the collector of the reference transistor, the two can be replaced by a single transistor with split-level input. Further applications will be shown in this form.

The GATE signal for the +AND is generated as a single line rather than as the $+N$ and $-P$ lines required in the 2-level system. This reduces the cost of the gate powering logic and the uncertainty as to the actual time of gate opening. Using a single-input +AND to translate the voltage level in the feedback path permits either the in-phase or inverted output of the +OR to supply the feedback path, depending on the loading of the trigger.

Data Latch

The 3-level technique also simplifies the data-latching function considerably. Fig. 6 shows two third-level circuits performing the function of at least four in the 2-level version.

When the $+N$ HOLD DATA line is false, its current into the supervisory input of the +AND insures that neither output from that block will conduct. In this state the +OR functions normally in every respect. The inverted output will carry current if there is an input current from either $+P$ DATA INPUT A or $+P$ DATA INPUT B and, conversely, the in-phase output will conduct if neither input is true.

When the $+N$ HOLD DATA lines go positive, current in the +AND is free to flow in accordance with the $+N$ DATA input. If at that time the +N DATA input is false and conducting, the out-of-phase output will conduct and apply the current to the split-level input of the +OR. Because the split-level input is more positive than the A and B inputs, the $+N$ DATA line will continue to conduct regardless of any change in the status of the input lines. The outputs will remain false as long as the $+N$ HOLD DATA line remains true.

If, at the time the $+N$ HOLD DATA input goes true, the +OR is satisfied, the $+N$ DATA input will cause the in-phase output of the +AND to conduct. The in-phase output is fed back as a normal-level input to the +OR, where it reinforces the data inputs. If both data inputs should go false, the +OR will remain satisfied. The outputs will remain true as long as the $+N$ HOLD DATA line remains true.

It is therefore apparent that both logical ones and zeros are held or latched in the +OR under control of the +AND. The circuit provides a mechanism for stabilizing a transient-input signal. The phasing of data movement must be controlled to insure that the +OR is not switching when the holding signal is applied. The exception to this rule is when the latch is being used as a synchronizing device, a case discussed in detail later.

As illustrated in Fig. 6, the circuit is performing a +OR of input-data signals. A −AND operation can be performed equally well by the same configuration. This is in contrast to the 2-level latch where the logical function must be an OR. Because of this flexibility the third-level latch can always be integrated with combinational logic so that no serial delay is added to a transfer path.

The third-level latch can provide either the in-phase or inverted-data output as a current line to feed a coaxial transmission line, special coupling network, etc. This flexibility results from the ability of the latching +AND to accept either polarity of data output as the normal-input signal.

Only a single holding signal is required by the third-level latch. This reduces the cost of powering logic and facilitates accurate control of the time of latching. The cost of the latching function does not exceed six transistors per bit regardless of the number of input signals. As with the trigger function it is frequently practical to wire OR input signals.

Forcing Circuit

The third-level latch and trigger which have been described have a common circuit configuration. This is redrawn in Fig. 7 and will be further considered because of its several interesting properties. As illustrated, this basic logical circuit is combinational in nature; the state of the input signals uniquely determines the state of the output signals. It is, however, necessary only to feed back appropriately the output signals to obtain the previously discussed latch and trigger functions and other useful circuits.

The forcing circuit responds to three input current signals, identified as $(-N)X$, $(-N)C$, and $(+P)Y$, and generates complementary output current signals $(-N)Z$ and $(-N)\bar{Z}$. The outputs are related to the inputs by the equations

$$Z = CY \vee \bar{C}X \vee XY \tag{1}$$

$$\bar{Z} = C\bar{Y} \vee \bar{C}\bar{X} \vee \bar{X}\bar{Y} \tag{2}$$

where the middle terms $\bar{C}X$ and $\bar{C}\bar{X}$ define the mutually exclusive secondary signals a and b respectively. It will be clear that equation 1 defines the true state for both outputs of the +OR when it is recalled that the in-phase output is in the positive or no-current condition when the OR is satisfied. Equation 2 is merely a simplified negation of equation 1 to define the false output condition.

The equations may be interpreted as follows: The output current Z will be present if input current X is present and input current C is absent. It will also be present if input current Y is present and the current combination of C absent and X absent does not exist. This is merely to say that if current b is present the split-level input will override input Y so that Z cannot be true.

It can thus be seen that, in the presence of a current at the third-level input C, the +OR block responds to input signal(s) Y as though the +AND circuit did not exist. If, on the other hand, the current C is not present the outputs of the +OR block are completely determined by the status of input signal(s) X. This is the unique forcing property of the circuit which results from the particular specification of third-level inputs in combination with the cross-coupling of both outputs of the +AND into the +OR.

A trigger function with bipolar gating is derived from the combinational forcing circuit by feeding the Z output back as the Y input after conversion of signal level. The defining equations for this sequential circuit then become

$$Z=(z \vee a)\bar{b} \qquad z=Z \tag{3}$$

$$A=\bar{C}X \qquad a=A \tag{4}$$

$$B=\bar{C}\bar{X} \qquad b=B \tag{5}$$

where, in accordance with conventional notation[6] the lower-case secondary signals represent outputs available only after time delay. In a similar manner the feedback of the Z output as the X input will generate the data-latching function. In this configuration the defining equations are:

$$Z=(Y \vee a)\bar{b} \qquad z=Z \tag{6}$$

$$A=\bar{C}z \qquad a=A \tag{7}$$

$$B=\bar{C}z \qquad b=B \tag{8}$$

It will be noticed that the data latch has effectively two feedback paths, while the trigger has only one.

LATCH-GATED TRIGGER

The addition of a second +OR block in parallel with that of the basic forcing circuit makes it possible to combine the latch and trigger functions. The latch-gated trigger is shown in Fig. 8. This is a single sequential circuit which itself provides both the storage and the latching functions to insure data stability during the period of setting. Economy is achieved in this circuit by the +AND performing both the gating and the holding functions.

There is no essential difference between the operation of the latch-gated trigger and that of a data latch feeding a trigger with bipolar gating. An important advantage, however, is a complete absence of phase difference between the latching and gating signals, because in this configuration they are identical. The operation may be briefly summarized as follows. When the $+N$ GATE signal is false the input +OR responds normally to the $+P$ DATA INPUT signals(s) and the trigger loop formed by the final +OR and +AND maintains a stable storage of either a true or false bit. If, when the $+N$ GATE goes true, the $+P$ DATA INPUT is present yielding the $-N$ BIT, the $+P$ SET ON line will be generated. This signal performs two functions: It is fed back into the input +OR where it guarantees that the $-N$ BIT will remain even if the $+P$ DATA INPUT should go false, and it switches the +OR of the trigger if the trigger is "off." When the +OR is switched the in-phase output goes to $+N$ ON, which switches the +AND resulting in $+P$ ON being fed back to latch the trigger "on."

When the $+N$ GATE goes true with the $+P$ DATA INPUT and $-N$ BIT absent, the $+P$ SET OFF signal is generated. This signal is fed back as a split-level input to the +OR where it insures that the $-N$ BIT cannot be true even if +P DATA INPUT does become true. The same line as a split-level input to the +OR of the trigger overrides the $+P$ ON and causes the $\pm N$ ON signals to become false. In turn the false $+N$ ON causes the $\pm P$ ON lines to go false, latching the trigger off.

During the gating period, data that have been impinging on the gating +AND are stabilized by the latch portion of the circuit, enabled forward, and set into the trigger element. At the same time new input data are blocked from penetrating the +OR. Conversely, during the nongating period, stable data are offered to succeeding stages of logic by the trigger portion of the circuit while the input +OR is receptive to new data. A cycle of operation is thus defined by the gating signal. It must be present sufficiently long to insure that the trigger is set and it must be absent long enough to allow data penetration. The over-all length of a cycle will be determined by the amount of external logic through which data must pass. Within these limits the waveshape used for gating may vary considerably. It has been found convenient to use the easily generated symmetrical waveform. The operation of the latch-gated trigger is analogous to that of a mechanical escapement.

REGISTER POSITIONS

As with the 2-level version a third-level trigger must be provided with additional functions to obtain a complete machine-register position. Fig. 9 shows a typical configuration having two input gates (one of the latch type), three true output gates, and two complement output gates. The output gates are of the distribution type described above. Hence only one true gate and one complement gate can be present at the same time. This is a small limitation because in normal operation a register will receive information from a single source and send it to a single destination in any particular cycle. This register would approximately replace the current-switching trigger and gating of Fig. 2 together with the data latch of Fig. 3.

The operation of the register may be inferred from the descriptions given for the third-level gated trigger and latch-gated trigger. The interesting modification of the trigger involved in the register is the combination of output gating and feedback functions. It was seen in the trigger circuit that a translated ON signal had to be fed back to hold the trigger on. In the register position the third-level $-N$ ON

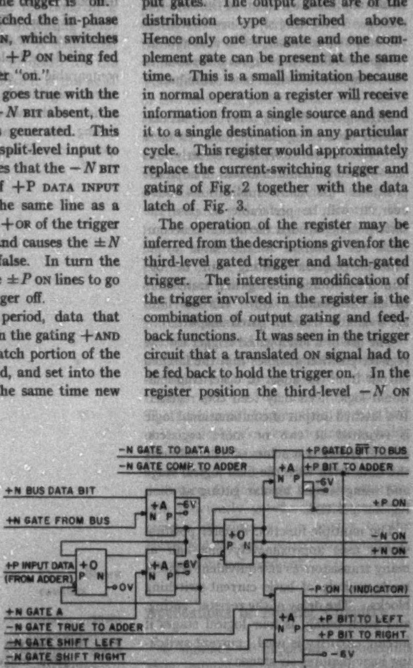


Fig. 9. Register position

signal, which enables the complement-gating distribution AND, yields an output which is translated and inverted. This $+P$ ON line is then available for feedback. In a similar manner a $-P$ ON signal is obtained from the transistor receiving the $+N$ ON input to the true-gating AND. Because indication is usually provided at a remote location, an output current from each register position must be provided. The $-P$ ON line is suitable for this purpose. The in-phase outputs of the distribution ANDS have an output of the form [(bit) (no true gate)]. This function is seldom useful with the result that the reference transistors can be replaced by diodes.

The register position as illustrated would be suitable for the principal accumulator of a machine. The names of the gates suggest typical machine applications. Data are gated into an external adder in either true or complement form. Because of the multiple functions of the distribution ANDS this is acccomplished at a cost of only two transistors per bit. The input OR of the latch gate is usually the final OR stage of the adder. Parallel read-in and read-out of data words may be accomplished by means of the data-bus gating. The gating of the inverted bit to the output bus makes use of a double inversion to allow more equal utilization of the complement and true distribution ANDS.

High-speed shifting may be accomplished by feeding the $+P$ BIT TO LEFT and $+P$ BIT TO RIGHT lines directly into the $+P$ INPUT DATA terminals of adjacent positions. In many applications, however, it will be preferable to pass the shifted data through the adder to accomplish the addition in the same cycle. Multiple-position shifting can be obtained by adding more gating transistors to the distribution AND. Current lines are available from all gated outputs.

The use of a latch-gated trigger for the register position assumes that the latched output from the adder or other combinational logic need only feed this register. If a latched output of combinational logic is required at two or more registers, greater economy is achieved by combining the data latch with the final logical stage and using simple bipolar gating at each register position.

The multiple-function register position shown uses approximately one half as many transistors as an equivalent register made entirely of basic current-switching blocks. The delay of the register, gating, and latch ORing, is four logical stages if forcing circuitry is used; current-switching logic requires six stages.

Forcing Circuitry as Sequential Logic

The forcing circuitry was developed by empirical techniques to solve specific problems in the design of high-performance computers. It is instructive, however, to study the circuits using the techniques of sequential circuit analysis. A better understanding of their operation and limitations can be obtained by this means. This method can also serve to clarify the differences between the various individual circuits.

ANALYSIS OF LATCH-GATED TRIGGER

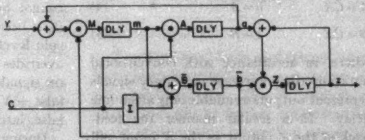
The most elaborate and complex of the forcing circuits that have been discussed is the latch-gated trigger. Its operation was described in current logical terms, without a systematic cataloging of states and transitions. One of the principal advantages of this circuit is its elimination of race conditions while using a minimum of components. Although the absence of races may be clear from the description, the actual transitions will now be analyzed.

In Fig. 10 the circuit of Fig. 8 is first redrawn in elemental form to show the feedback loops and to specify the locations of the effective circuit delays. There is a delay associated with each output of a logical block and the delays are comparable for all blocks. The I block merely signifies logical inversion. Comparison of the two drawings will show that among the secondary signals m corresponds to $-N$ BIT, a to $+P$ SET ON, b to $+P$ SET OFF and z to $+P$ ON. The inputs Y and C, are respectively, the $+P$ DATA INPUT and $+N$ GATE signals. The equation for each of these signals is written below the circuit diagram.

Fig. 10 also shows a truth table for all possible states of the circuit. From this the stable states can be selected and the transitions that will occur can be predicted. This is done in the chart, Fig. 10, in which the circled numbers indicate stable states and the uncircled numbers define states that must (by the change of a single secondary signal) switch without race to the corresponding stable state. Each table entry with a dot represents a state that, if entered, would lead to a critical race. The race transitions for one set of these states are shown as dashed lines. In these charts changes in input or primary variables cause movement between columns; changes in secondary variables cause transition between rows within a column.

The behavior of the circuit can be analyzed from the chart by examining the transitions which occur when a stable state is acted upon by an input-data change. The chart shows, for example, that from stable-state 3 transition through unstable states 1000 and 1100 to state 4 follows the gate C going true. Alternatively, the data going false cause a transition from state 3 back to state 2 through state 1000.

Examining the possible transitions from each of the eight stable states shows that each leads to a new stable state either



$M=(Y\vee a)\bar{b}=m$ $A=mC=a$ $B=\bar{m}C=b$ $Z=(z\vee a)\bar{b}=z$

	YC			
mabz	01	00	10	11
0000	0010	0000	1000	1010
0001	0011	0001	1001	1011
0010	0010	0000	0000	0010
0011	0010	0000	0000	0010
0100	1011	1001	1001	1011
0101	1011	1001	1001	1011
0110	0010	0000	0000	0010
0111	0010	0000	0000	0010
1000	0100	0000	1000	1100
1001	0101	0001	1001	1101
1010	0100	0000	0000	0100
1011	0100	0000	0000	0100
1100	1101	1001	1001	1101
1101	1101	1001	1001	1101
1110	0100	0000	0000	0100
1111	0100	0000	0000	0100
	MABZ			

YC			
01	00	10	11
1	②	3	•
1	⑥	7	•
①	2	3	⓪
1	•	•	0
•	•	•	•
•	6	7	•
1	•	•	0
1	•	•	0
•	2	③	4
•	6	⑦	4
•	2	3	•
•	•	•	•
5	•	•	4
⑤	6	7	④
•	•	•	•
•	•	•	•

Fig. 10. Analysis of latch-gated trigger

directly or by completely defined secondary changes. In no case does the change of a single input variable lead to a race condition. It is, of course, assumed that the circuit is allowed sufficient time to stabilize itself between input changes.

Although only one known application allows simultaneous change of input signals, consider what happens in such an event. Fig. 10 shows that a transition from a stable state in which the gating signal is present to one in which it is absent always leads unambiguously to another stable state. On the other hand, a transition from a stable state in which the gate is absent to one in which it is present accompanied by a change in input data always leads to a critical race condition. Thus, if from state 3 the gate goes true and data false, transition is to state 1000 in column 01. Here two secondary signals must change simultaneously and the resulting race can lead to either of states 1 and 5. A worse possibility is that state 1000 may change to state 0100, which may lead back to 1000 and cause oscillation.

The possible race conditions might be anticipated by an analysis of the delays in the circuit. The latching is accomplished by one of two signals fed back from the gating AND to the input OR. The selection of the feedback signal depends on whether the output of the input OR is true or false at the instant of gating. If the output of the OR changes just as the feedback is generated, because of a simultaneous data change, reversal in feedback output could occur. The first feedback could again change the OR output and lead to oscillation.

In practice it has been found that a "simultaneous" change in inputs to the latching AND means that the data-input signal is crossing reference as the control input releases the block current, which then divides between the two outputs. The split-level network develops a swing equal to the normal input so that the current in the OR block also divides. This maintains the data input to the AND near reference and causes a finite delay before the circuit latches true or false. The maximum observed delay for a decision to be latched is less than 100 nsec and there is no oscillation.

Distinctions Between Forcing Circuits

The three classes of forcing circuit (trigger, data latch, latch-gated trigger) that have been presented are distinguished by more than circuit variation. They perform different logical functions in a machine and represent increasing complexity and logical power. A sequential analysis was presented in detail for the latch-gated trigger because it is the most complex and interesting. Only the results of such investigations on the others will be mentioned here.

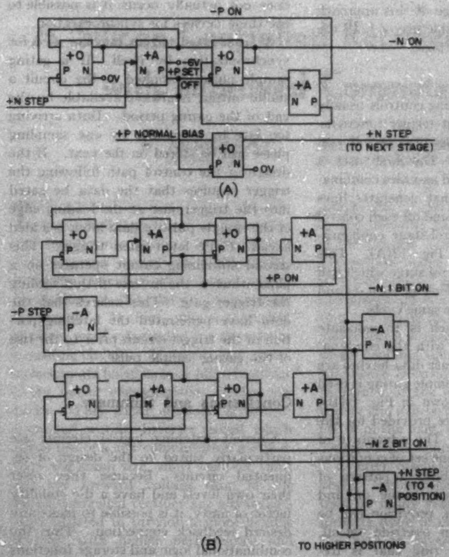


Fig. 11. Binary counters

A—Serial count propagation

B—Parallel count propagation

The biopolar gated trigger is the most familiar of the circuits. It has a single feedback path. When a gating signal is applied new information flows through the trigger and is presented to further logic. A change in information during the gating period can pass directly through the trigger. When the gating signal drops, the trigger stores the state of the information during the gating period. No change in information can propagate through the trigger at this time. Ambiguity in switching can result only from changes of both data and gating which are close together but not coincident, or from too-short gating pulses. Gating flexibility makes the trigger advantageous for principal storage.

Although its usual application causes the data latch to appear functionally different from the trigger, the two are very nearly equivalent. The six stable states of the circuits are identical, and if the absence of a holding input is considered as a gating input the description given for trigger also applies to the latch. An important distinction, however, is that, during the gating period, data need pass through only one logical stage in the latch, ~~while two stages are~~ required in the trigger. This difference is related to the presence of two feedback paths in the latch. It is also reflected in a critical race condition resulting from simultaneous data change and latching. The data latch's ability to merge with combinational logic makes it attractive for use as secondary storage.

The latch-gated trigger performs the functions of two cascaded triggers operating on inverted gate signals. The latch portion of the circuit is receptive to new data when the gate is not present. At this time the trigger portion is holding stored information. When the gate is present the latch portion of the circuit stores the information which reached it during the nongating period, while the trigger receives the new information. The latch-gated trigger then, in one half of its cycle, blocks new information, holds the data it has, and gates it on; in the other half of its cycle it receives new data and stores for following logic the information previously gated. The latch-gated register provides a d-c analog to the Haven's delay unit, permitting simultaneous read-in and read-out.

Applications of Forcing Circuitry

A number of important applications were implied by the described register

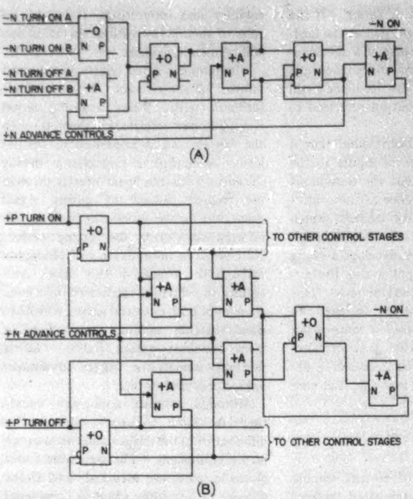


Fig. 12. Control stages

A—Individually generated switching lines

B—Switching lines driving multiple triggers

position. These include simple storage and accumulation, and shifting registers. All of the functions usually performed by sequential logic can be generated from the basic trigger and latch circuits. A few of the more common are presented here.

Binary Counter

A binary counter with serial-carry propagation can be constructed by trivial additions to the latch-gated trigger. As shown in Fig. 11(A) the $+P$ ON output of the trigger is fed back as the data input. This yields a binary trigger action that is itself important for control applications. After level conversion the $+P$ SET OFF signal provides the stepping pulse to the next stage of the counter. It is doubtful, however, that a counter of this type would be used. The propagation of the count is delayed by two logical stages for each bit of the counter. This delay quickly accumulates to the point where equal performance can be achieved with circuits that are less costly.

Fig. 11(B) shows the first two stages of a counter having parallel count propagation. In this case the stepping pulse for each state is generated by an ANDing of the stepping control with the "on" state of all lower-order bits. As the size of a counter of this type is increased, power is required for the control line and lower-order bits; propagating AND circuits must be cascaded to avoid exceeding the allowable number of inputs to a block.

It has been found convenient in many cases to have the stepping control directly drive the gates of all counter bits. On an advance, each counter position is either inverted or again set to the same state. In this form the counting logic becomes distinct from the register positions and assumes the characteristics of an add-one adder. The advantage of this approach is that it permits flexible use of both the register and counting logic.

Control Stages

Sequencing and gating controls usually consist of independent trigger functions. Two general approaches may be followed in organizing these. The first uses a latch-gated trigger and assumes combinational control logic that generates lines specifically to turn on and off each control trigger. Such a control-stage configuration is illustrated in Fig. 12(A). This approach yields a control stage which will be turned on if conditions for turn-on and turn-off are met on the same cycle.

The second approach is to associate the latching function with the combinational logic. In this case data latches are used in the logic and simple gating is used on the triggers, as shown in Fig. 12(B). Separate AND gates are provided for the on and off function. The advantage of this approach is that one set of conditions can more easily establish a pattern of control triggers, turning some on and others off. As shown, provision can be made to have set-on override set-off.

The fixed-sequence ring that is frequently employed in controls can be implemented most readily in forcing circuitry by the use of a shifting register similar to that of Fig. 9. The multidirectional gating out of the register is not required. An alternative approach is to use the binary counter and decode the individual states.

Synchronization

Synchronization of communication between units operating at different frequencies is a complex control problem in the design of any computer system. Signals from a slower unit must be sampled by the timing of a faster unit, and vice versa. The pulse to be synchronized must be at least as long as one sampling cycle to insure that the communication is not missed. To accomplish synchronization a storage device must accept a data input arriving at a random time with respect to the sample. For no phase relationship may it generate short "sliver" responses or commence oscillation. However, the output of the synchronizing circuit need not be predictable for all possible phase relationships.

With direct-coupled logic it is not a simple matter to design a circuit that will guarantee that the synchronization requirements are met. The latching circuits that have been presented here all have a theoretical potential for oscillation when phasing of data and sample is not controlled. Because the oscillation does not actually occur, it is possible to use these devices for synchronization.

A latch-gated trigger may be used for synchronization by itself. If a gating sample pulse is applied to the circuit a stable output is always available by the end of the gating period. Data arriving too late to be caught by one sampling phase will be stored on the next. If the design of the control path following the trigger requires that the data be gated into the trigger just at the leading edge of the sample pulse, a data latch is added ahead of the latch-gated trigger. This second stabilizing circuit operates on a pulse which is the inverse of that feeding the trigger gate. This insures that the data have penetrated the latching portion of the trigger circuit prior to the rise of the gating sample pulse.

Conclusions and Summary

Current-switching logical circuits are particularly suited to the design of sequential circuits. Because they reset their own levels and have a d-c stability factor of unity, it is possible to make any desired feedback connection. Thus the combinational logic and storage functions may be integrated. Machine performance is correspondingly increased through both cost reduction and faster operation. These same characteristics also make current-switching circuits important in buffering between stages of faster but degenerative circuits.

The current-switching circuits can be modified to increase greatly their power as sequential elements. The first essential modification is introduction of a third-input signal level; the second is a double cross connection of two logical blocks to permit the status of one block to be either completely determined by, or completely independent of, the status of the other block. The status of a control input to the second block determines the mode of operation of the first. Appropriate choice of input and feedback terminals on this basic forcing circuit will generate either a set–reset trigger with bipolar gating or a data-latching function.

These circuits can be interconnected to perform all the storage functions that are required for machine operation. The use of a latching function which becomes part of combinational logic eliminates the setting time of a second register from the machine cycle. The use of a latch-gated trigger with distribution gating centralizes the storage and routing functions of a machine and makes them independent of the remainder of the logic. This configuration represents a universal register which can be used in any application.

Sequential analysis of the forcing circuits aids in understanding their operational limitations. All are free of critical race conditions as long as the inputs are changed individually and sufficient time is allowed for each state to stabilize. It becomes clear that the differences between the circuits are more in terms of function and delay rather than sequential characteristics.

The techniques of sequential-circuit synthesis could be used to design the entire control logic for a machine. In practice, however, the savings in design time that results from using a standard set of sequential building-block functions will usually more than compensate for any increase in component count. All of the usual control functions are readily generated using these forcing circuits.

Packaging considerations also dictate the use of a standard set of sequential circuits. In order to minimize the number of configurations which must be packaged, circuits must be selected to have the greatest versatility and widest usage. A standard IBM double card[2] will hold one large (multiple input and output gating) register position, two smaller registers, or three data latches.

The small voltage swings required as inputs and the well-defined current outputs make these circuits compatible with many other types of logical blocks. The circuits are tolerant of signal noise and logical skew. Close phase control is not required. The presence of stable voltage levels is an aid to machine servicing. The use of unipolar gating inputs having standard voltage swings leads to additional economy in powering logic. The high speed, logical power, and reliability of the forcing circuits make them a practical, inexpensive system of sequential logic.

References

1. Millimicrosecond Transistor Current Switching Circuits, H. S. Yourke. *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-4, no. 3, Sept. 1957, pp. 236–40.

2. The Engineering Design of the Stretch Computer, Erich Bloch. "Proceedings of the Eastern Joint Computer Conference," 1959, pp. 48–58 (available from AIEE).

3. The Applications of Transistors to Computers, R. A. Henle, J. L. Walsh. *Proceedings*, Institute of Radio Engineers, vol. 46, 1958, pp. 1240–54.

4. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components (book), J. von Neumann. "Automata Studies," Princeton University Press, Princeton, N. J., 1956, pp. 43–98.

5. Switching Circuits and Logical Design (book), S. H. Caldwell. John Wiley & Sons, Inc., New York, N. Y., 1958, pp. 453–513.

A Silicon-Controlled Rectifier Inverter with Improved Commutation

W. McMURRAY
MEMBER AIEE

D. P. SHATTUCK
ASSOCIATE MEMBER AIEE

THE ANNOUNCEMENT of the silicon controlled rectifier[1,2] in December 1957 instituted a new era in the field of electric power conversion. Almost immediately, attention was focused on the application of this device in inversion equipment.[3] At first, silicon-controlled rectifiers were substituted for thyratrons in existing circuits. However, such direct replacements failed to take full advantage of the better characteristics of the new device. This paper presents a silicon-controlled rectifier inverter circuit which makes use of a new commutation method. Here the small size, short turn-off time, and low forward voltage drop of the silicon-controlled rectifier are advantageously utilized to provide a more flexible inverter with higher performance and smaller size. The circuit was developed in the General Engineering Laboratory of the General Electric Company.

Inverters using the improved commutation method have many advantages, including high efficiency and low regulation. A single-phase version of the inverter has a square-wave output voltage which remains square over the complete load range, from no load to full load. The frequency of the inverter can be varied widely and it is possible to vary the power factor of the load, without changes in the commutation components.

A brief discussion of one of the inverter arrangements which have been used for many years will be presented to indicate some of the historic problems in the field of inversion. Following this, the operation of the new inverter circuit will be described. An analysis of the technique will be presented, with experimental confirmation. Practical considerations associated with the inverter are also treated.

Background

A brief consideration of one of the inverter approaches most frequently used in the past will permit a convenient look into the background of the field of inversion. The paralleled capacitor commutated inverter, see Fig. 1, commonly called the parallel inverter, is one of several circuits which have been successfully applied for many years. Much has been written about the parallel inverter, but two papers by C. F. Wagner[4,5] are of particular significance. The parallel inverter is analyzed with resistive loads in

---

Paper 61-718, recommended by the AIEE Semiconductor Rectifiers Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted December 2, 1960; made available for printing May 2, 1961.

W. McMurray and D. P. Shattuck are with the General Electric Company, Schenectady, N. Y.

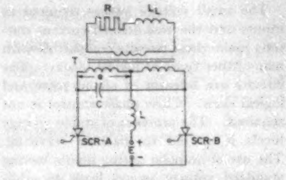


Fig. 1. Parallel inverter. All three windings on the transformer T have the same number of turns

the first paper and with inductive loads in the second paper. A summary of the information presented in these two papers will help to illustrate a few of the problems which are associated with this inverter approach.

As the author of these papers indicated, the circuits considered were intended "to operate with any of the grid-controlled mercury-vapor discharge tubes or the conventional grid-controlled rectifiers." The silicon-controlled rectifier also meets the device requirements, i.e., "current conduction may be initiated at any instant when the anode-cathode voltage has proper polarity, but current extinction must be accomplished through some function of the circuit."

An important characteristic calculated by Wagner is shown in Fig. 2. It relates the time t_0 available for turning off the controlled rectifiers, expressed as a fraction of the period T of the a-c output, to the parameter K for several values of power factor PF, where

$$K=4(2\pi f)CZ \tag{1}$$

$$Z=\sqrt{R^2+(2\pi fL_L)^2} \tag{2}$$

$$PF=R/Z \tag{3}$$

This characteristic can be used to determine the value of commutating capacitance C required with the minimum value of Z (maximum load condition) in order to insure an adequate turn-off time. It is apparent that the short turn-off time of the silicon-controlled rectifier enables a small capacitor to be used, especially when the load is purely resistive; see Fig. 2, curve for $PF=1.0$.

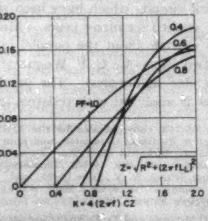


Fig. 2. Curves for turn-off time

A look at the load voltage waveforms with resistive load (Fig. 3) indicates how the peak voltage goes up as the value of K increases. The peak value of voltage approaches infinity as the load resistance approaches infinity (no-load). Obviously, a minimum load must be maintained to limit the peak voltage, which is impressed on the controlled rectifiers. In addition to the limitations on load range, starting is difficult since no-load starting is not possible.

The characteristics of Fig. 2 and 3 result when the d-c ballast inductance L is assumed infinite. With a practical sized inductance the turn-off time characteristic (Fig. 2) is relatively unchanged while the voltage waveform becomes oscillatory, as shown in Fig. 4. Purely resistive loading is still assumed. The parameter J is defined as the ratio of the inductive reactance of L to the load resistance:

$$J=2\pi fL/R \tag{4}$$

It can be seen that the overshoot in voltage is much greater for low values of J and high values of K. Again the sensitivity to load appears, since J decreases and K increases with increasing values of R.

Further difficulties arise when inductive loads are applied to the parallel inverter (Fig. 1 with $L_L>0$). From Fig. 2 it can be seen that, regardless of how short the turn-off time of the controlled rectifier may be, the value of C must be quite large in order to provide any turn-off time at all, with an inductive load. Again an increase in the peak load voltage, which is also applied to the controlled rectifiers, occurs as the load decreases (K increases). This is illustrated in Fig. 5 for a load power factor of 0.8, again assuming infinite ballast inductance. At lower power factors the problem becomes even more acute, because the required value of K increases, as seen in Fig. 2. To accommodate a variable power factor load, the value of K and consequently the value of C must be fixed at the maximum required value. This results in high peak voltages when the load changes to a higher power factor. Switching the values of C as the load changes in magnitude or power factor is possible but in most cases impractical.

To sum up, it can be said that the parallel inverter in its simplest form, as considered by C. F. Wagner, has several shortcomings: a limited load range, need for a large commutating capacitor when inductive loads are applied, and the requirement of a rather large inductance to maintain continuous direct current and limit voltage overshoots. Furthermore, starting the inverter is difficult since no-load operation is not possible. The starting problem exists because of the long time required to charge the large commutating capacitors. The voltage attained during the first half-cycle may not be sufficient to insure adequate turn-off time for the first commutation.

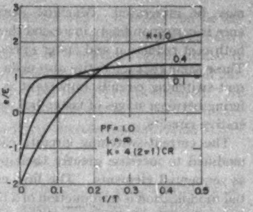


Fig. 3. Voltage relations, infinite ballast inductance (resistive load)

While the problems discussed pertain primarily to the parallel inverter, they are representative of difficulties associated with applying silicon-controlled rectifiers directly in any of the commonly used inverter circuits of the past. The inherent advantages of the silicon-controlled rectifier are not fully utilized in such direct replacements.

Improved Mode of Commutation

Six circuit arrangements for single-phase inverters embodying the improved mode of commutation are shown in Fig. 6. With a transformer turns ratio of unity, each circuit generates a nearly square wave of magnitude E and can commutate load currents up to a maximum magnitude I_m. The required values of capacitance and inductance in each case are related to the values C and L, respectively, required by arrangement (A). In (C) and (D) the two halves of the primary winding of the output transformer are assumed to be tightly coupled, which in practice may require special design

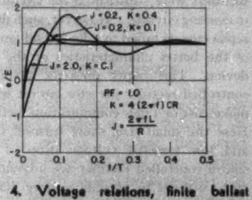


Fig. 4. Voltage relations, finite ballast inductance (resistive load)

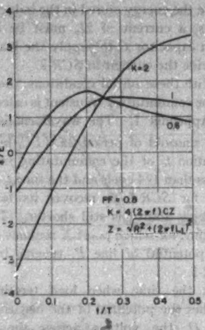


Fig. 5. Voltage relations, infinite ballast inductance (inductive load)

techniques such as interleaving or bifilar winding. In the other circuits, close coupling between the two halves of the center-tapped inductance is necessary.

Circuits (A) and (B) require a center-tapped d-c supply, or a neutral point established by a pair of large capacitors, which can be of the electrolytic type. Arrangement (C) corresponds to the conventional parallel inverter previously described, modified to utilize the improved method of commutation. Of course, the conventional parallel inverter can also exist in configurations similar to the other arrangements shown in Fig. 6. The bridge circuits (E) and (F) are biphase versions of (A) and (B). The inductances in each half of the bridge may or may not be coupled. In circuit (F), when the inductances are not coupled, the two halves of the bridge may be operated out of phase, enabling the load voltage to be controlled by pulse width modulation. Circuits (A) and (B) are also basic building blocks for polyphase bridge inverters.

A common feature of the improved inverters is the use of rectifiers to return or "feed back" to the d-c supply the reactive power associated with inductive or capacitive loads. In fact, comparing circuit (C) with the conventional parallel inverter of Fig. 1, these rectifiers are the only difference evident in the circuit diagram. With a capacitive load, for example, some of the energy drawn from the d-c supply through a controlled rectifier during the first part of a half-cycle of load voltage is temporarily stored in the load capacitance, and then fed back to the d-c supply through a "feedback" rectifier later in the same half cycle. In the case of inductive loads, energy stored in the load inductance during the latter part of a half-cycle is returned to the d-c supply during the first part of the next half-cycle. The feedback rectifiers are connected to taps on the primary winding of the output transformer, rather than to the ends of the winding, in order to enable the feedback of energy stored and otherwise "trapped" in the inductance L just after commutation, as will be described later. This results in some variation of the output voltage with the power factor of the load, but far less than is the case with the conventional parallel inverter.

By handling reactive loads in this manner, the need for a large commutating capacitor and a large d-c choke or "ballast inductance" is eliminated, since these elements no longer have to store large amounts of reactive energy. Instead, a much smaller commutating capacitor C can be used in conjunction with a small inductance L, which will be termed the "commutating inductance." The natural period $2\pi\sqrt{LC}$ is directly related to the turn-off time of the controlled rectifiers, so that the short turn-off time of silicon-controlled rectifiers is fully exploited. The characteristic impedance $\sqrt{L/C}$ is related to the d-c supply voltage E and the maximum value I_{oo} of load current which must be commutated. Selection of the commutation parameters L and C is discussed in Appendix IV.

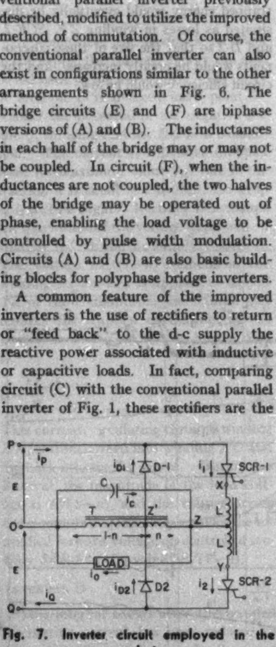
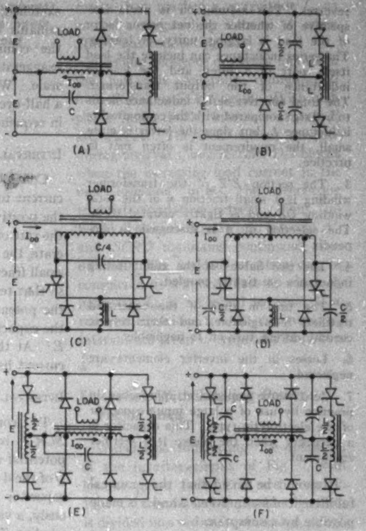


Fig. 7. Inverter circuit employed in the analysis

Fig. 6 (right). Inverter circuits with the improved mode of commutation



Theory of Operation

For the purposes of explanation and analysis, the circuit arrangement of Fig. 7 is selected. This is the circuit of Fig. 6 (A) in which the load impedance (including any filter elements) and the impedance of the output transformer, both referred to the primary winding, are lumped together as the total effective load on the inverter. An ideal autotransformer T is assumed in order to establish a tap point Z' for connection of the feedback rectifiers. To simplify the analysis, the following idealizing assumptions are made:

1. The period of the a-c output of the inverter is very much greater than the turn-off time of the silicon-controlled rectifiers.
2. Series inductance included as part of the total load will maintain the load current substantially constant at a value I_o throughout the commutating interval, defined as the interval during which the load voltage

reverses. This assumption is made irrespective of whether the net power factor of the load is lagging, unity, or leading. The series inductance can include the load itself, part of a filter, and the leakage inductance of the output transformer. The total effective series inductance needs to be large compared with the commutating inductance L, but since the latter is quite small, the requirement is often met in practice.

3. The section $Z'Z$ of the transformer winding is a small fraction n of the total winding OZ and is tightly coupled to it. The selection of n is discussed in Appendix V.
4. The two halves of the commutating inductance are tightly coupled.
5. The turn-on time of the controlled rectifiers is negligible, and their reverse current during turn-off is negligible.
6. Losses in the inverter elements are neglected.
7. Zero source impedance, approached in practice by use of a large input capacitor of the electrolytic type. This capacitor is a necessity when the source is a rectifier equipment.

It should be noted that the practical fulfillment of assumptions 2 and 3 is made possible by assumption 1.

Operation With Lagging Power Factor Load

One half-cycle of operation is illustrated in the waveforms sketched in Fig. 8. The voltages of points P Q, X, Y, Z, Z' as indicated in Fig. 7 are shown with respect to the d-c neutral point 0. Across any circuit element, the voltage is the the difference between the appropriate curves. For example, the voltage across controlled rectifier *SCR-1* is the difference in voltage level between P and X. The locations of the currents sketched in Fig. 8 are also specified in Fig. 7.

Each half-cycle is divided into five intervals, designated A, B, C, D, E, which are bounded by changes in the state of a controlled rectifier or a feedback rectifier from a nonconducting to a conducting condition or vice versa. Such a change effectively alters the equivalent circuit of the inverter. The operation can be completely analyzed by writing the equations for each equivalent circuit and matching the boundary conditions. However, this approach would be unduly complex and depend upon the load impedance so that general conclusions would be difficult to extract.

From the standpoint of the designer, the most important part of the cycle is the commutating interval, B in Fig. 8. Since it is so short, it is drawn with an expanded time scale. An arbitrary waveform for the load current i_o is sketched, and the approximation that i_o remains constant at a value I_o during the commutating interval B appears exaggerated due to the expanded time scale. With reference to Figs. 7 and 8, a half-cycle of operation will be described in sequence.

Interval A

Controlled rectifier *SCR-1* is conducting current to the load, drawing power from the positive line P of the d-c supply. If the rate of change of load current is moderate, the voltage across inductance L is a small fraction of the d-c supply voltage E so that terminal Z of the load is close to the potential of the positive line P and the capacitor C is charged to the voltage E. At the end of interval A, the load current has attained a value I_o.

Interval B

To commutate, controlled rectifier *SCR-2* is gated "on," dropping the potential of point Y to the negative line Q of the d-c supply. Since the voltage on capacitor C cannot change instantaneously, a voltage of $2E$ must appear across winding ZY of inductance L. The same voltage is induced in winding XZ, placing an inverse voltage of $2E$ on controlled rectifier *SCR-1* and turning it off. The current I_o originally flowing in *SCR-1* and winding XZ must be transferred to winding ZY and *SCR-2* in order to preserve the energy stored in the inductance. Thus, a current of $2I_o$ must be drawn from capacitor C to supply the load and provide the current in *SCR-2*.

With these initial conditions, the form of the commutating transient is calculated in Appendix I. The waveform is part of a sinusoid of period $2\pi\sqrt{LC}$, but the duration t_c of the commutating interval is less than 1/4 cycle and the time t_o available for *SCR-1* to recover its forward blocking ability is still shorter. This is the time for which point X remains above the potential of line P, reverse biasing *SCR-1*.

At the time when load terminal Z reaches the potential of the negative d-c line Q, the voltage across the commutating inductance is zero and the current i_2 has built up to a maximum value I_m. If the feedback rectifiers were connected to terminal Z instead of the tap Z', *D-2* would now conduct and the current I_m would circulate through winding ZY, *SCR-2* and *D-2*, thereby "trapping" the energy $1/2\, LI_m{}^2$. Feedback rectifier *D-2* would also carry the load current, feeding energy from the load back to the negative side of the d-c supply, until such time as the load current reversed. However, the trapped energy would eventually be dissipated as losses in winding ZY, *SCR-2*, and *D-2* so that the size of these components would be greatly increased, and the circuit efficiency would be low.

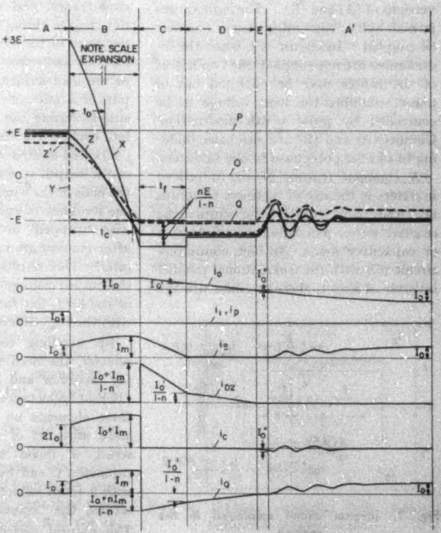


Fig. 8. Waveforms with lagging power factor load

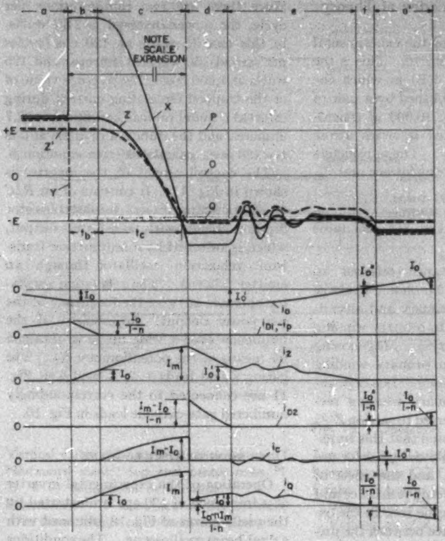


Fig. 9. Waveforms with leading power factor load

INTERVAL C

By connecting the feedback rectifiers to the tap Z' the trapped energy can also be fed back to the d-c supply. The onset of interval C is delayed until point Z' reaches the potential of the negative d-c line Q, bringing rectifier *D-2* into conduction. The current i_2 will have dropped from its maximum value I_m, but if the fraction n is small (0.1 to 0.2), the drop in i_2 will be slight. Rectifier *D-2* clamps the magnitude of the voltage impressed on winding OZ' to the d-c supply voltage E. The voltage $nE/(1-n)$ induced in the winding section $Z'Z$ is applied to section YZ of the commutating inductance and reduces the trapped circulating current linearly from the value I_m to zero in time t_f.

$$t_f=\frac{LI_m(1-n)}{E\qquad n}\tag{5}$$

This current, circulating through winding $Z'Z$, is transformed into winding $Z'O$ and fed back into the d-c source. During this interval, the magnitude of the load voltage is $E/(1-n)$, and the voltage across the commutating inductance raises the reapplied forward voltage on controlled rectifier *SCR-1* to the value $2E/(1-n)$.

INTERVAL D

This interval occurs after the trapped circulating current has been reduced to zero, but the continuing feedback of energy from the inductive load keeps rectifier *D-2* in conduction. The load voltage remains at the value $E/(1-n)$. The voltage $nE/(1-n)$ induced in winding section $Z'Z$ now appears as inverse bias across *SCR-2*.

INTERVAL E

When the load current reverses its polarity, rectifier *D-2* blocks and the increasing load current in the new direction is supplied by discharge of the capacitor C until the potential of terminal Z reaches the level of the d-c line Q.

INTERVAL A'

Controlled rectifier *SCR-2* now commences to conduct again In order to do so after being reverse biased during intervals D and E, a positive gating signal must be reapplied at this time, or else the gating voltage applied at the start of interval B must have been maintained. It is found more convenient to use square-wave gating signals.

Since the load current built up during interval E cannot immediately pass through the inductance L, the capacitor C initially continues to discharge, starting a period of "ringing" between L and C. The peak ringing voltage is less than $nE/(1-n)$. Eventually, the ringing is damped out by the load and by losses. At the end of the interval A', the load current has reached a magnitude I_o, as in the corresponding interval A when *SCR-1* was conducting.

A similar sequence of events follows in the half-cycle following commutation from *SCR-2* to *SCR-1* again. With high power factor loads, intervals D and E are absent. The load current reverses during interval C and rectifier *D-2* blocks when the increasing load current in the new direction meets the decreasing trapped circulating current. Thus, interval C merges directly into interval A', and *SCR-2* maintains continuous conduction. The no-load condition can be considered as a special case in which i_o, including I_o, is set equal to zero. The reduced equations for the commutating interval are given in Appendix II of this paper.

Operation With Leading Power Factor Load

Waveforms illustrating the operation of the inverter circuit of Fig. 7 with leading power factor load are drawn in Fig. 9. One half-cycle is shown, which is divided into six intervals, *a, b, c, d, e, f*. The commutating intervals *b* and *c* have an expanded time scale. An arbitrary load current waveform is sketched, and during intervals *b* and *c* an approximately constant magnitude I_o is assumed. The sequence of operation in a half-cycle will now be analyzed, with reference to Figs. 7 and 9.

INTERVAL A

The load current is negative and rectifier *D-1* is conducting, feeding energy from the load back into the positive side of the d-c supply. The voltage across the load and capacitor C is raised to the value $E/(1-n)$ by the action of the autotransformer T. Controlled rectifier *SCR-1* is inversely biased by the voltage $nE/(1-n)$ induced in winding $Z'Z$. At the end of interval *a* the load current is $-I_o$ and the current in *D-1* is $I_o/(1-n)$.

INTERVAL B

Commutation is initiated by gating "on" controlled rectifier *SCR-2*, effectively connecting point Y to the negative line Q of the d-c supply. The capacitor C maintains point Z at its former potential, so that a voltage of $E(2-n)/(1-n)$ is impressed across winding ZY of the commutating inductance. The same voltage induced in winding XZ increases the inverse voltage on *SCR-1* to $2E/(1-n)$. However, in the case of leading loads, this turn-off action is redundant.

The current i_2 increases linearly from zero to the value I_o in a time t_b given by

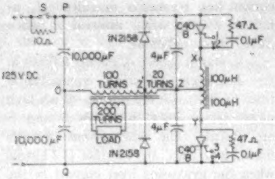


Fig. 10. Experimental power circuit

$$t_b = \frac{LI_o(1-n)}{E(2-n)} \qquad (6)$$

During the same period the current in *D-1* is reduced to zero, after which *D-1* blocks. Essentially, the load current has been transferred from the path through *D-1* to the path through *SCR-2*.

INTERVAL C

The capacitor *C* now enters into the commutating transient, increasing the current i_2 from I_o to a peak value I_m as it reverses its charge. Interval *c* is analyzed in Appendix III and lasts for a little over 1/4 cycle of the natural frequency of *L* and *C*.

INTERVAL D

This interval is quite similar to interval *C* with lagging load. Rectifier *D-2* conducts and the tapped transformer permits recovery of the excess energy trapped in the inductance *L* after commutation. The current i_2 falls linearly from the peak value I_m at the rate of $nE/(1-n)L$ amperes per second until i_2 is equal to the load current, assumed to have reached a value I_o' by the end of interval *d*.

INTERVAL E

Rectifier *D-2* blocks, and the magnitude of the load voltage falls from $E/(1-n)$ to approximately *E*, after a period of "ringing" between capacitor *C* and inductance *L*. This is similar to interval *A'* with lagging load.

INTERVAL F

When the load current reverses, controlled rectifier *SCR-2* blocks, and the increasing load current in the new direction charges capacitor *C* back to the level $E/(1-n)$ volts.

INTERVAL A'

Rectifier *D-2* starts conducting again, and feeds energy from the load back into the negative side of the d-c supply. The conditions are now reversed from the original state in interval *a*.

Experimental Verification of Theory

The power circuit of the experimental inverter is shown in Fig. 10. This is the configuration of Fig. 6(B) in which the neutral point *0* is established by a pair of electrolytic capacitors, 10,000 μf (microfarads) each, connected in series across the 125-volt d-c supply. The parameters of the power output transformer are:

Secondary winding, 200 turns.
Primary winding *OZ*, 120 turns.
Primary winding section *Z'Z*, 20 turns ($n = 1/6$).

Total winding resistance referred to primary winding *OZ*, 0.6 ohm. Leakage inductance between primary and secondary windings, referred to primary winding *OZ*, 600 μh (microhenrys). Magnetizing inductance referred to primary winding *OZ*, 113 mh (millihenrys). Leakage inductance between primary winding sections *OZ'* and *Z'Z*, referred to section *Z'Z*, 4 μh. It should be noted that this transformer was not orginally designed for use in this inverter circuit, and the effect of the relatively large winding resistance and leakage inductance can be seen in the oscillograms as a difference between the primary and secondary voltage waveforms.

The values of commutating capacitance *C* (8 μf) and inductance *L* (100 μh) required to commutate a maximum load current I_{oo} of 17.7 amperes referred to the primary winding *OZ*, with controlled rectifiers having a turn-off time t_{oo} of 12 μsec (microseconds), were calculated from equations 40 and 41. The two halves of the commutating inductance were wound bifilar on a molybdenum permalloy powder core. With resistive load, the maximum power output theoretically obtainable is about 62.5 volts×17.7 amperes, or 1,106 watts. Under no-load conditions, the energy *W* trapped in the choke after commutation is:

$$W = 4.7EI_{oo}t_{oo} = 4.7\times62.5\times17.7\times12\times10^{-6} = 0.0624 \text{ joule} \qquad (7)$$

Since there are two commutations per cycle, the power involved is $2Wf$ watts, in this case 15 watts at 120 cps (cycles per second), 50 watts at 400 cps, and 125 watts at 1,000 cps. The average value of of the trapped circulating current during interval *C* under no-load conditions is 17.7 amperes, and the duration of interval *C* is $t_f = 283$ μsec, calculated from equation 5.

The control circuit of the inverter is shown in Fig. 11. It consists of an *R/C* (resistance-capacitance) coupled transistor flip-flop circuit, with transformer output, which is switched by a unijunction transistor relaxation oscillator through an emitter follower. The relaxation oscillator operates at twice the frequency of the a-c power output. Adjustment of the frequency over a wide range is obtained by means of the potentiometer R_T. The gating signal leads 1, 2, 3, and 4 in Fig. 11 are connected to the correspondingly numbered gate-cathode leads in Fig. 10.

DISCUSSION OF OSCILLOGRAMS

Operation of the experimental inverter at a frequency of 120 cps is illustrated by the oscillograms of Fig. 12, obtained with a dual-beam oscilloscope. The conditions of loading are:

1. No load. The magnetizing impedance of the output transformer actually imposes a light load of very low, lagging power factor on the inverter.
2. Resistance load of 12 ohms. The impedance of the tranformer results in an effective load on the inverter of 4.9 ohms in series with 0.6 mh, referred to the primary winding *OZ*. The high power factor load eliminates intervals *D* and *E* from the waveforms, as discussed in the theory of operation. The transformer leakage inductance is insufficient to prevent a 40% drop in load current during the commutating interval, so that the assumption of constant load current during this period is not strictly valid. Examination of the lower oscillogram

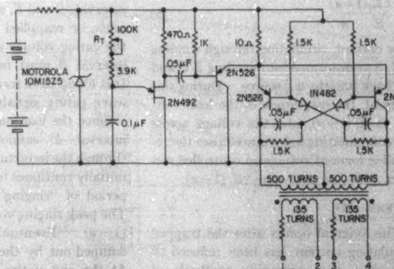


Fig. 11. Experimental control circuit

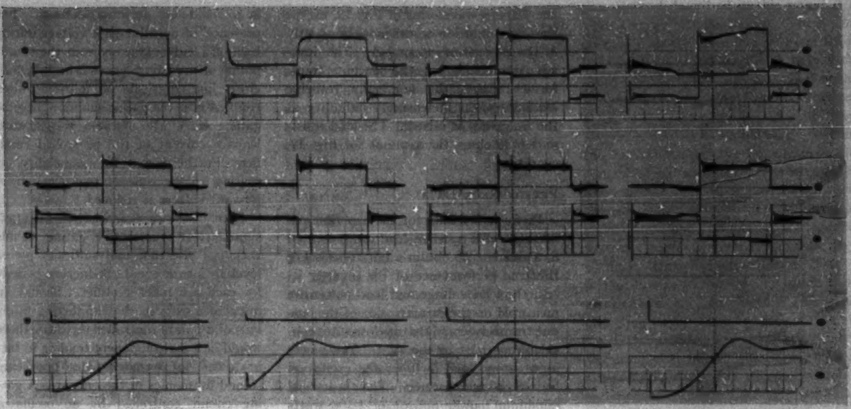


Fig. 12. Operation of Inverter at 120 cps

Vertical scale: 100 volts per division
Horizontal scale: Top and center rows, 1 millisecond per division
Bottom row: 10 microseconds per division, to show details of commutating interval

Top row: load voltage (upper traces)
Primary voltage OZ (lower traces)
Center and bottom rows:
Voltage on SCR-2 (upper traces)
Voltage on SCR-1 (lower traces)

First column: no load
Second column: resistive load
Third column: lagging load
Fourth column: leading load

showing detail of the commutating interval reveals that the available turn-off time is reduced to about 12 μsec; the condition therefore represents the maximum safe level of loading.

3. Lagging load consisting of 12 ohms in series with 12.5 mh. Including the transformer impedance, the total load referred to the primary winding is 4.9 ohms plus 5.1 mh. With a sinusoidal voltage at 120 cps, the power factor would be close to 0.8. The time constant of about 1/1,000 second is short enough compared with the half-cycle of 1/240 second so that the load current at commutation is the same as in condition 2, however, the series inductance is sufficient to prevent appreciable change in load current during the commutation interval. Nevertheless, little difference can be seen between the oscillograms of the commutating intervals for conditions 2 and 3.

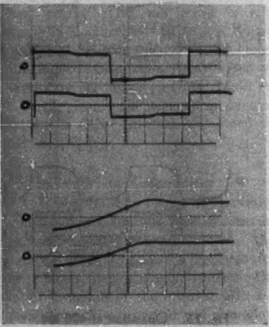


Fig. 13. Voltage OZ (upper traces) Voltage OZ′ (lower traces)

Vertical scale: 100 volts per division
Horizontal scale: top, 1 millisecond per division
Bottom: 10 microseconds per division

4. Leading load consisting of 50 μf in series with 12.5 mh plus some resistance, resulting in a moderate load of net leading power factor.

The experimental waveforms verify the theory of operation which has been developed, for both lagging and leading power factor loads. All of the intervals described in the theory and drawn in Figs. 8 and 9 can be distinguished in the oscillograms of Fig. 12. One feature of the oscillograms which is not explained by the theory is the slightly underdamped oscillatory overshoot in voltage which occurs just after the commutating interval. This is due to the leakage inductance between primary winding sections OZ' and $Z'Z$, as confirmed by the fact that the overshoot appears on the voltage from O to Z, but not on the voltage from O to Z'; see Fig. 13. The frequency of oscillation is close to the natural frequency of the effective commutating capacitance, 8 μf, and the leakage inductance referred to section $Z'Z$, 4 μh. The overshoot is undesirable because it increases the peak forward voltage applied to the controlled rectifiers by about 25%.

Another feature not covered in the theory is the large "spike" of inverse voltage which appears on the controlled rectifiers at the beginning of the commutating interval. The duration is so short that the spike is not visible in the waveforms of Fig. 12. Expanding the time scale, the transient appears as shown in Fig. 14. Its cause is the considerable reverse current which is required to turn off the controlled rectifiers. This current is established in the lead inductance and the leakage inductance of the commutating reactor, and when the controlled rectifier suddenly recovers its inverse blocking ability, "ringing" occurs with the capacitance of the controlled rectifier, plus any stray capacitance.

Without any suppression, the oscillation frequency is about 8 megacycles, and the peak voltage attained is 2.5 times the ideal peak inverse voltage applied to the controlled rectifiers. This is for the no-load condition, shown in Fig. 14(A).

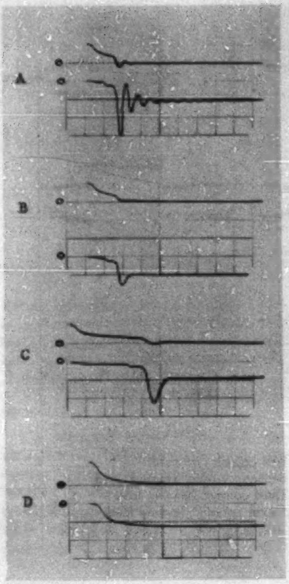


Fig. 14 Turn-off transient
Voltage on SCR-2: upper traces
Voltage on SCR-1: lower traces

A—No load, without suppression
B—No load, with suppression
C—Maximum load, lagging power factor
D—Leading power factor load
Vertical scale: 100 volts per division
Horizontal scale: 0.2 microsecond per division

With the addition of an *R/C* filter connected directly across the terminals of the controlled rectifier, as shown in Fig. 10, the peak overshoot is considerably reduced, see Fig. 14(B). The resistance is selected to produce critically damped, or slightly underdamped, conditions. Such a filter also aids in suppressing externally produced voltage transients on the controlled rectifiers.

With loading, the forward current, and hence the reverse recovery current, of the controlled rectifiers increases. This raises the overshoot voltage, as shown in Fig. 14(C). However, when the load has a leading power factor, the controlled rectifiers are turned off before the commutating interval, and no overshoot appears; see Fig. 14(D).

Operation of the inverter at 400 cps is shown in Fig. 15 and at 1,000 cps in Fig. 16. The waveforms are very similar to the 120-cps waveforms of Fig. 12, except that commutation occurs more frequently at the expense of interval *A*. The leakage inductance of the output transformer makes the nominally resistive loading appear increasingly inductive as the frequency is raised. The load resistance is 12 ohms, the same as for Fig. 12, condition 2.

PERFORMANCE DATA

Performance data for the experimental inverter with resistive load are presented in Table I. To obtain a fair appraisal of the true performance of the inverter itself, the load voltage and load power are measured on the primary side of the output transformer, so that its losses are considered as part of the load. The power input at no load includes the excitation loss of the transformer. The theoretical maximum power output of 1,106 watts is not achieved with an adequate margin of turn-off time, for several reasons. First, the leakage inductance of the transformer prevents true resistive loading, especially at the higher frequencies. The magnetizing current of the transformer adds to the load current at the time of commutation. Then, losses during the commutation interval cause a reduction in the actual time for which the controlled rectifiers are reverse biased.

From Table I, it is seen that the inverter efficiency is high at 120 cps and 400 cps. The lower efficiency at 1,000 cps is partly due to the limitation of load power by the high impedance transformer, as well as the more frequent commutation loss. Voltage regulation from no load to maximum load is approximately 12%. Most of the drop occurs between no load and 20% load. This is because of the light, low *PF* conditions which actually prevail at no load, so that the d-c supply voltage is impressed across the section *OZ'* of the transformer winding for the major part of a cycle. Relatively light resistive loading produces considerable improvement in power factor, and the supply voltage is now applied to the total winding *OZ* for the major part of a cycle.

Practical Considerations

GATING CIRCUIT

The gating signals applied to the silicon-controlled rectifiers must be square waves, rather than pulses, in order to operate properly with inductive loads, as explained in the theory. While the square-wave gate signal is positive when the controlled rectifier is to conduct, it is negative when the controlled rectifier is to hold off forward voltage. The presence of negative gate voltage during hold-off is advantageous from the standpoint of avoiding false firing due to either line transients or spurious pulses coupled into the gating circuit. Inverse gate voltage also reduces the forward leakage current of the controlled rectifier. Furthermore, the possibility of refiring at the time of reapplication of forward voltage is reduced.

Care must be exercised to maintain symmetrical square-wave gating signals with fast rise time and sufficient power level to insure equal conduction periods for each controlled rectifier. Should an unbalance exist in the conducting intervals of a pair of controlled rectifiers, there would be a d-c component tending to saturate the transformer besides producing second harmonics in the output. A gating circuit controlled by an oscillator operating at twice the output frequency, such as shown in Fig. 11, will produce very little unbalance between alternate half-cycles.

WIRING LAYOUT

Lead dress is a very important consideration in this inverter, as in others. There are two major factors to consider when laying out the wiring of an inverter of this type. First, care must be taken to avoid any excessive lead inductance, which could prevent proper circuit performance. A good rule is to maintain small spacing between the wires car-

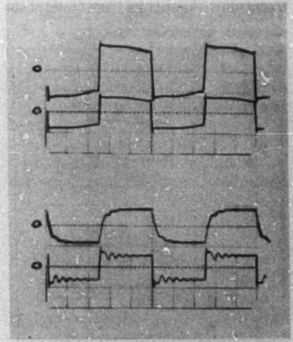


Fig. 15. Operation at 400 cps

Load voltage: upper traces
Primary voltage OZ: lower traces
No load: top
Resistive load: bottom
Vertical scale: 100 volts per division
Horizontal scale: 500 microseconds per division

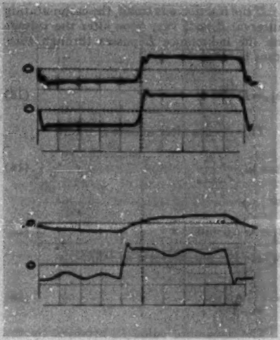


Fig. 16. Operation at 1,000 cps

Load voltage: upper traces
Primary voltage OZ: lower traces
No load: top
Resistive load: bottom
Vertical scale: 100 volts per division
Horizontal scale: 100 microseconds per division

rying heavy current to and from any particular portion of the equipment. Second, the layout should be planned to minimize any coupling between power and control circuits. Inverters of this type have rapidly changing currents in the power circuit. Such rapid current transient tend to becoupled into the gating circuit if no precautions are taken.

Both problems can be minimized by the use of proper lead dress. The inverter components should be laid out to minimize lead lengths. This will result in low wiring inductance as well as low wiring resistance and consequently reduce losses and improve regulation. Any loops in wires carrying heavy currents should be avoided. Leads in the gating circuits should be either shielded or twisted to reduce coupling.

INPUT FILTER

It is important to have a dynamically stiff d-c source for this type of inverter to insure proper commutation. If the source is not stiff because of line or source impedances, it is necessary to shunt the input terminals of the inverter with capacitance. The capacitor will also allow for pump back of current when the inverter is reactively loaded should the d-c source be unable to accept reverse current, as in the case of rectifier equipments. In this event the capacitor must be sufficiently large to avoid excessive voltage fluctuation.

CAPACITOR SELECTION

High current pulses which flow in the commutating capacitor require care in the selection of the type of capacitor to be used. The rms value and not the average value of current in the capacitor must be considered. The use of extended foil construction has proved very helpful in limiting heating.

STARTING

~~With the improved~~ inverter circuit, the starting problem is greatly reduced. The small commutating capacitance can be charged to full voltage in a short time and the inverter can start and operate under no-load conditions. However, precautions must still be taken to prevent saturation of the transformer during starting. If a small air gap is provided in the core to minimize any residual flux, starting may be accomplished by one of two means. The core can be designed to operate at less than half the saturation value of flux density, which in high-frequency inverters may be necessary anyhow to prevent excessive core loss. Or, the gating circuit can be arranged to halve the duration of the first half-cycle.

Where an input filter capacitor is employed, the initial charging transient from an inductive source can produce a peak voltage of up to twice the steady-state d-c supply voltage. This increases the possibility of transformer saturation besides applying an overvoltage to the controlled rectifiers. All of these starting problems are easily solved by using the procedure employed with the experimental inverter. The control circuit is energized first, applying the gating signals to the controlled rectifiers. The main power switch is then closed, with the switch *S* (Fig. 10) open so that a starting resistance is placed in series with the d-c supply. Thus the voltage on the input filter capacitor builds up relatively slowly with no overshoot, allowing adjustment of the transformer flux to near steady-state a-c conditions before full voltage is attained. The starting resistance is short-circuited by closing switch *S* before applying load to the inverter. An added advantage of this method is that in the event that some malfunction develops during starting, resulting in a shoot-through, the starting resistance will limit the current and prevent damage to the inverter.

Conclusions

One of the most important features of inverters with the improved commutation method described in this paper is a high degree of flexibility. They can operate over the complete range of loads from no load to full load with very little change in the peak voltage applied to the silicon-controlled rectifiers. Furthermore, the power factor of the load may be varied over a wide range without the need for altering the value of the commutating capacitor. The inverter is also capable of operating over a wide frequency range. Both the commutating capacitance and inductance are smaller than in previous inverter circuits, particularly with inductive loads. This is primarily due to the way in which these components are employed. The commutating capacitance and inductance are tuned to produce a commutating interval somewhat longer than the required turn-off time of the silicon-controlled rectifiers under the condition of peak current which must be extinguished. As a result, the load power factor only affects the values of the capacitance and inductance to the extent that it determines this peak current.

The efficiency of the inverter is high and regulation is low because of the low forward drop of the silicon-controlled rectifier and low losses in the small d-c inductance. The elimination of high peak voltages on the controlled rectifiers enables units of lower voltage rating to be used, and can eliminate, or reduce the number of series connected units in high voltage circuits. The voltage rating of the transformer is also reduced, since the voltage rise at light load is eliminated.

The inverter is particularly advantageous in applications where the inherent square-wave output is required or can be tolerated. The full savings in reduced component size are realized in such applications. In cases where sinusoidal output is required, the savings in commutating component size are lost because inductance and capacitance added for filtering raise the total component size to a level comparable with that of other inverter approaches. Many applications, however, require overcurrent protection of the

Table I. Performance of Inverter with Resistive Load

Output frequency, cps	120	400	1,000
Primary voltage OZ, rms volts:			
No load	74.5	75.0	78.0
Maximum load	66.0	66.2	66.9
Maximum load power, watts	655	585	300
Efficiency at maximum load, %	96	92	67
Power input at no load, watts	18	35	64

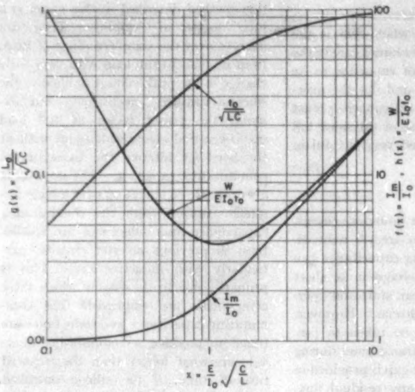


Fig. 17. Commutation parameters of inverter

inverter as well as sinusoidal output. In these applications current limiting can be accomplished by the same components which filter the output. Here the need for filtering does not detract from the size and weight advantages.

The new method of commutation results in considerable performance improvement, compared with previous arrangements such as the conventional parallel inverter of Fig. 1. A reduction in size is also achieved. For example, to accommodate the third load condition illustrated in Fig. 12 ($Z = 6.23$ ohms, $PF = 0.8$), the parallel inverter of Fig. 1 would require $K = 0.4$ (see Fig. 2) and $C = 21.3\ \mu f$. With less than infinite ballast inductance, C would be somewhat larger. The new circuit in an equivalent configuration, Fig. 6(C), requires a capacitor of only $2\ \mu f$. The key to the improved circuitry is the use of feedback rectifiers to handle reactive loads. Other inverters employing a technique similar to this have recently been developed,[6] and further improvements are foreseen in the future.

Appendix I. Analysis of Commutating Interval B, Lagging Power Factor Load

Referring to Figs. 7 and 8, time is measured from the instant when controlled rectifier *SCR-2* is turned "on." The current i_2 flowing in controlled rectifier *SCR-2* is chosen as the independent variable.

The initial conditions at time $t = +0$ are:

Current in inductance $L = i_2(+0) = I_0$
Voltage on capacitor $C = e_2(+0) = E$

The differential equation describing the commutating transient may be written as follows:

$$E = L\frac{di_2}{dt} - E + \int_0^t \frac{i_2 + I_0}{C}\,d\tau \qquad (8)$$

The Laplace transform of equation 8 is

$$\frac{2E}{s} = Lsi_2(s) - LI_0 + \frac{i_2(s)}{sC} + \frac{I_0}{s^2C} \qquad (9)$$

Solving equation 9 for $i_2(s)$, the transform of $i_2(t)$

$$i_2(s) = \frac{\dfrac{2E}{L} + sI_0 - \dfrac{I_0}{sLC}}{s^2 + \dfrac{1}{LC}} \qquad (10)$$

The inverse transform of equation 10 gives

$$i_2 = \frac{2E}{\omega L}\sin\omega t + I_0\,(2\cos\omega t - 1) \qquad (11)$$

where

$$\omega = \frac{1}{\sqrt{LC}} \qquad (12)$$

If the fraction n is small, the commutating interval B ends very soon after the voltage on the inductance L passes through zero, that is, when

$$e_{y2} = L\frac{di_2}{dt} = 0 \qquad (13)$$

This occurs after a time t_c where

$$\tan\omega t_c = \frac{E}{\omega L I_0} \qquad (14)$$

or

$$t_c = \sqrt{LC}\tan^{-1}x \qquad (15)$$

where the parameter x is defined as

$$x = \frac{E}{I_0}\sqrt{\frac{C}{L}} \qquad (16)$$

The maximum value I_m attained by the current i_2 is given by

$$I_m = 2\sqrt{\frac{C}{L}E^2 + I_0^2} - I_0 \qquad (17)$$

or

$$\frac{I_m}{I_0} = f(x) = 2\sqrt{x^2+1} - 1 \qquad (18)$$

The time t_0 during which controlled rectifier *SCR-1* is reverse biased is given by

$$\frac{t_0}{\sqrt{LC}} = g(x) = \sin^{-1}\frac{x}{\sqrt{x^2+1}} - \sin^{-1}\frac{x}{2\sqrt{x^2+1}} \qquad (19)$$

The energy W trapped in the inductance L after commutation is

$$W = \frac{1}{2}LI_m^2 = \frac{1}{2}LI_0^2[f(x)]^2 \qquad (20)$$

$$\frac{W}{EI_0t_0} = h(x) = \frac{[f(x)]^2}{2xg(x)} = \frac{(2\sqrt{x^2+1} - 1)^2}{2x\left(\sin^{-1}\dfrac{x}{\sqrt{x^2+1}} - \sin^{-1}\dfrac{x}{2\sqrt{x^2+1}}\right)} \qquad (21)$$

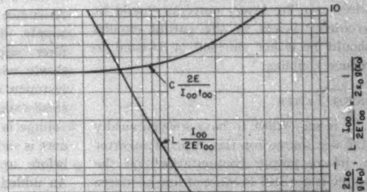


Fig. 18. Variation of C and L with parameter x_0

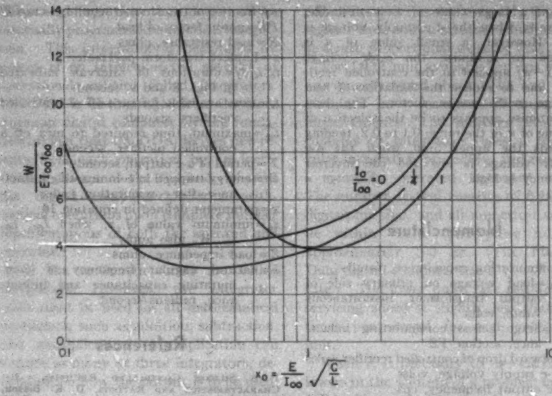


Fig. 19. Variation of trapped energy with parameter x_0 and load

The commutation parameters defined by equations 18, 19, and 21 are graphed in Fig. 17. The required values of C and L are determined by the d-c supply voltage $2E$, the maximum value I_{oo} of load current at the instant of commutation, and the maximum turn-off time t_{oo} of the controlled rectifiers (12 μsec for the General Electric *C40*).

In addition, the designer may select the minimum value x_o of the parameter x. The values of C and L are then calculated from the equations:

$$C\frac{2E}{I_{oo}t_{oo}}=\frac{2x_o}{g(x_o)} \tag{22}$$

$$L\frac{I_{oo}}{2Et_{oo}}=\frac{1}{2x_o g(x_o)} \tag{23}$$

The functions represented by equations 22 and 23 are plotted in Fig. 18.

Note that after determination of C and L, variation in x will be solely due to changes in load I_o. As I_o is reduced from the maximum value I_{oo}, x will increase from its minimum value x_o, and the actual time t_o available for turning off the controlled rectifiers will increase from its minimum value t_{oo}.

$$\frac{I_o}{I_{oo}}=\frac{x_o}{x} \tag{24}$$

Appendix II. No-Load Condition

The equations describing the commutating interval when the inverter is unloaded may be derived from the equations for the case of lagging power factor load by setting I_o equal to zero, or x infinite.

Equation 11 becomes

$$i_2=\frac{2E}{\omega L}\sin\omega t \tag{25}$$

and

$$t_c=\frac{\pi}{2}\sqrt{LC}=\frac{\pi t_{oo}}{2g(x_o)} \tag{26}$$

$$t_o=\frac{2}{3}t_c \tag{27}$$

$$I_m=2E\sqrt{\frac{C}{L}} \tag{28}$$

$$\frac{I_m}{I_{oo}}=2x_o \tag{29}$$

$$W=2CE^2 \tag{30}$$

$$\frac{W}{EI_{oo}t_{oo}}=\frac{2x_o}{g(x_o)} \tag{31}$$

Appendix III. Analysis of Commutating Interval c, Leading Power Factor Load

With reference to Figs. 7 and 9, time will be measured from the instant when rectifier *D-1* blocks.

The initial conditions at time $t=+0$ are:

Current in inductance $L=i_2(+0)=I_o$

Voltage on capacitor $C=e_{os}(+0)=E/(1-n)$

The differential equation describing the commutating transient in terms of i_2 as the independent variable may be written:

$$E=L\frac{di_2}{dt}-\frac{E}{1-n}+\int_0^t\frac{i_2-I_o}{C}\,d\tau \tag{32}$$

The Laplace transform of equation 32 is

$$\frac{E}{s}\left(\frac{2-n}{1-n}\right)=Lsi_2(s)-LI_o+\frac{i_2(s)}{sC}-\frac{I_o}{s^2C} \tag{33}$$

Solving equation 33 for $i_2(s)$, the transform of $i_2(t)$

$$i_2(s)=\frac{\frac{E}{L}\left(\frac{2-n}{1-n}\right)+sI_o+\frac{I_o}{sLC}}{s^2+\frac{1}{LC}} \tag{34}$$

Applying the inverse transformation to equation 34

$$i_2=\frac{E}{\omega L}\left(\frac{2-n}{1-n}\right)\sin\omega t+I_o \tag{35}$$

It follows that

$$t_c=\frac{\pi}{2\omega}=\frac{\pi}{2}\sqrt{LC} \tag{36}$$

and

$$I_m=I_o+E\left(\frac{2-n}{1-n}\right)\sqrt{\frac{C}{L}} \tag{37}$$

Appendix IV. Selection of the Commutation Parameters x_o, C, L

The major criterion for selecting the optimum commutation parameters is judged to be the minimization of the energy W trapped in the inductance L after commutation. This energy not only determines the rating of the inductance, but also increases the circuit losses because of imperfect efficiency in feeding the energy back to the d-c supply. It is seen from Fig. 17 that the selection $x_o=1.15$ will minimize the trapped energy W_o when commutating the maximum lagging load current I_{oo}.

$$\frac{W_o}{EI_{oo}t_{oo}}=h(x_o)=3.87\text{ minimum} \tag{38}$$

However, this choice of x_o does not minimize the trapped energy W when the load current at commutation I_o is less than the maximum value I_{oo}. The energy W can be related to the fixed quantity $EI_{oo}t_{oo}$ with the aid of equations 38, 20, and 24:

$$\frac{W}{EI_{oo}t_{oo}}=h(x_o)\frac{W}{W_o}=h(x_o)\left(\frac{I_of(x)}{I_{oo}f(x_o)}\right)^2$$

$$=h(x_o)\left(\frac{x_of(x)}{xf(x_o)}\right) \tag{39}$$

In Fig. 19, the variation of $W/E\cdot I_{oo}\cdot t_{oo}$ with x_o is shown for the following load conditions:

1. $I_o=I_{oo}$, $x=x_o$, maximum load, equation 38
2. $I_o=I_{oo}/4$, $x=4x_o$, calculated from equation 39
3. $I_o=0$, $x\rightarrow\infty$, no load, equation 31

It is seen that if x_o is within the range 0.75 to 1.15, the value of W changes little with load, and is close to a minimum. For most inverter designs, the selection of $x_o=1$ is judged to be the optimum, and the following results are obtained:

$$C=\frac{t_{oo}I_{oo}}{0.425E} \tag{40}$$

$$L=\frac{t_{oo}E}{0.425I_{oo}} \tag{41}$$

The peak current in a controlled rectifier is the maximum value of I_m, which occurs at no load or otherwise when $I_o=0$, and is equal to $2I_{oo}$, equation 29. The maximum value of trapped energy W also occurs under the same conditions, and from Fig. 19 is equal to $4.7\,EI_{oo}t_{oo}$. If the inverter operates with a leading power factor load, the maximum value of I_m is greater, according to equation 37 and the maximum value of W is correspondingly greater.

Another parameter of interest is the rate of rise of reapplied forward voltage on a controlled rectifier after it has been turned off. The steepest slope occurs when $I_o=I_{oo}$ and is nearly uniform. A forward voltage of $2E$ is reapplied in the time

$$t_{co}-t_{oo}=\left[\frac{\tan^{-1}x_o}{g(x_o)}-1\right]t_{oo} \qquad (42)$$

$$=0.85t_{oo} \qquad \text{when } x_o=1$$

The average rate of voltage rise is then $2E/0.85\ t_{oo}$. For the General Electric *C40* controlled rectifier, t_{oo} is 12 μsec and the maximum allowable rate of rise of voltage is 20 volts per μsec, which will not be exceeded provided that E is less than 102 volts.

Appendix V. Selection of the Tap Fraction n

The average current in the controlled rectifier during interval C is $I_m/2$. If a constant value e_f is assumed for the forward drop of the controlled rectifier, the energy dissipated in the *SCR* during interval C is

$$\frac{1}{2}I_m e_f t_f=\frac{LI_m^2 e_f(1-n)}{2En}=W\frac{e_f(1-n)}{En} \qquad (43)$$

where the value of t_f has been obtained from equation 5. Thus a certain fraction of the trapped energy W is lost in the controlled rectifier, and more is lost in the feedback rectifier and the windings of the inductance and the transformer.

For efficient recovery of the trapped energy, a large value of n is desirable, especially when the d-c supply voltage is low. However, a small value of n is desired to minimize the peak voltage $2E/(1-n)$ applied to the controlled rectifiers, and to reduce the variation of load voltage with power factor. The best compromise appears to be the selection of a value of n in the range 0.1 to 0.2, tending towards the upper limit when the d-c supply voltage is low and the inverter frequency is high.

Nomenclature

C=commutating capacitance, farads
e, e_{os}=load voltage on primary side of output transformer, instantaneous value, volts
e_{yz}=voltage across commutating inductance, section YZ
e_f=forward drop of controlled rectifier, volts
E=d-c supply voltage, volts
f=a-c output frequency, cps
$f(x)$, $g(x)$, $h(x)$=functions of parameter x, equations 18, 19, and 21
i subscripts 1, 2, $D1$, $D2$, c, o, P, Q=currents in inverter circuit, as indicated in Fig. 7, instantaneous values, amperes
I_m=peak current in controlled rectifiers, amperes
I_o=load current at time of commutation, amperes
I_{oo}=maximum value of load current at time of commutation, amperes
I_o', I_o''=load current at particular instants, see Figs. 8 and 9
J, K=parameters defined in equations 4 and 1
L=commutating inductance, henrys
L_L=load inductance, henrys
n=fraction of transformer winding to tap Z'
PF=power factor of load
R=load resistance, ohms
s=Laplace transform parameter
t_b, t_c, t_f=durations of intervals indicated in Figs. 8 and 9, seconds
t_o=time available for turn-off of controlled rectifiers, seconds
t_{oo}=maximum time required to turn off a controlled rectifier, seconds
T=period of a-c output, seconds
W=energy trapped in commutating inductance after commutation, joules
x=parameter defined in equation 16
x_o=minimum value of x, when I_{oo}, t_{oo}, t_{co}, W_o also apply
Z=load impedance, ohms
ω=natural angular frequency of commutating capacitance and inductance, radians/second

References

1. A SILICON CONTROLLED RECTIFIER; I—CHARACTERISTICS AND RATINGS, **D. K. Bisson, R. F. Dyer.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, May 1959, pp. 102–06.

2. CONTROLLED RECTIFIER MANUAL (book). *E.C.G.-442*, General Electric Company, Schenectady, N. Y., 1960.

3. SOLID-STATE POWER INVERSION TECHNIQUES, **D. A. Paynter, B. D. Bedford, J. D. Harnden, Jr.** *Semiconductor Products*, New York, N. Y. Mar. 1960, pp. 51–56; Apr., pp. 50–55.

4. PARALLEL INVERTER WITH RESISTIVE LOAD, **C. F. Wagner.** *AIEE Transactions*, vol. 54, Nov 1935, pp. 1227–35.

5. PARALLEL INVERTER WITH INDUCTIVE LOAD, **C. F. Wagner.** *Ibid.*, vol. 55, Sept. 1936, pp. 970–80.

6. STATIC INVERTER DELIVERS REGULATED 3-PHASE POWER, **M. Lilienstein.** *Electronics*, New York, N. Y., July 8, 1960, pp. 55–59.

A Comparison of Computers

F. G. CURL
NONMEMBER AIEE

A ROUTINE SIMULATING the operation of an analog computer was constructed for use on a digital computer at the Jet Propulsion Laboratory. This code, called DEPI,[1] as an abbreviation for "differential equations pseudocode interpreter," provides much of the flexibility and ease of operation associated with analog computer operation. A study was made in which the speed, accuracy, and flexibility of the DEPI system were compared with those attributes in a digital differential analyzer, an analog computer, and a digital computer. The vehicle of comparison was a zero-lift, zero-angle-of-attack missile-trajectory problem. The results of this comparison are presented here in tabular and graphical form. Because of variations which exist among commercial models of each type of computer used in the study, the salient features of the particular models are described.

Types of Computers Used in Comparisons

ANALOG COMPUTER

The analog computer used in this study is a system largely made up of components manufactured by Electronic Associates, Inc., as described in a previously published paper.[2] The components of the system are of two general types: electronic components (operational amplifiers and diode function generators) and electromechanical devices (servomechanisms and amplifier-operated relays). The operational amplifiers used for summing and integrating are the more accurate components. The servo devices used for multiplying, dividing, trigonometric resolution, function generation, and recording, together with the relays, are among the more inaccurate components.

A drum-type function generator was used to simulate atmospheric pressure in the comparison problem given in the Appendix. The solution is extremely sensitive to errors in atmospheric pressure. The generation of this function, and the effects of variations in it due to error, are further discussed in a following section. A thorough discussion of the electromechanical servomultiplier as a source of error is presented in reference 3.

DIGITAL DIFFERENTIAL ANALYZER

The digital differential analyzer (DDA) is a device for obtaining the solution of

---

Paper 61-725, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 8, 1961; made available for printing April 12, 1961.

F. G. CURL is with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, Calif.

differential equations by means of an internally programmed numerical integration. This integration is performed by means of an interconnected network of operational entities called integrators. The DDA is distinguished from the more common digital computer by the fundamental fact that the DDA derives new results from each operation by an incremental change of the operand, whereas the digital computer performs arithmetical operations upon complete number representations of both the operand and operator.[4]

All the arithmetic units are called integrators. However, the integrator units must be used for all mathematical operations, such as addition, subtraction, and multiplication. Multiplication can require as many as three integrators, depending on how the output is to be used. Such use of integrator units must be considered in comparisons of the numbers of integrators available in analog computers and DDA machines. Physically, a DDA contains only one arithmetic system, as do most digital computers, and the integrator units are actually memory subdivisions. Consequently, DDA operations are serial in nature, as are those of the digital computer.

In the simplest sense, each integrator consists of two arithmetic registers and a unit to transfer information from one to the other. These are designated as the *Y*-register, the add-unit, and the *R*-register.

Each integrator performs a repeated cycle of operations. In any particular cycle, the *Y*-register and/or the add-unit may receive incremental inputs ΔY and ΔX, respectively, and the *R*-register may emit an incremental output ΔZ. The *Y*-register initially set at Y_0 is an accumulating register and, during the nth cycle, contains

$$Y_n = Y_0 + \sum_{i=1}^{n} (\Delta Y)_i$$

While within the nth cycle, the *R*-register has added to its previous contents the quantity

$$\left[Y_0 + \sum_{i=1}^{n} (\Delta Y)_i \Delta X \right]$$

where $\Delta X = 0$ or 1; consequently, the contents of the *R*-register represent the sum of a series of rectangles whose bases are +1 (the smallest incremental step), and whose heights are the appropriate Y_n values. The *R*-register continues to receive and accumulate the successive contributions from the *Y*-register until an overflow occurs, creating an increment ΔZ. Successive overflows of the *R*-register then emit a stream of increments at a rate proportional to $\Delta Z = Y\Delta X$.

The DDA unit available for this study was manufactured about 1949, and later machines have incorporated important improvements. The machine available contains 60 integrators, uses a rectangular integration procedure, and requires a manually operated keyboard input. The number system consists of 6 binary digits plus sign, and all numerical values lie within ±1.0. The pulse rate is approximately 100 kc. (The Electro-Data digital computer has a 144-kc pulse rate.) The lower pulse rate results in a servicing speed of slightly less than 100 times per second for each integrator unit.

When it is necessary to use curve followers in the solution of a problem on the DDA, the speed of computation is greatly reduced, because the computer must wait until it has received point-to-point information before it can go on with the solution. This is analogous to the time required for tape read-in as a factor in computation speed on a digital computer. Accuracy and speed are reciprocal: for each binary-digit gain in accuracy, the solution time doubles.

Other machines of more recent manufacture incorporate one or more of the following advantages: doubling of the servicing-cycle speed when the total number of integrators required is less than one-half the maximum number available; perforated-tape input; trapezoidal integration, resulting in improved accuracy for each service cycle; and the use of decimal-number systems. Consequently, the machine available does not represent the DDA to the best advantage in performance comparisons with the analog computer and general-purpose digital machines.

DIGITAL COMPUTER

The digital computer which was used in this study is the Electro-Data computer described in detail in reference 1. This machine provides 4,080 10-decimal-digit words, with 4,000 main-memory words and 80 words of high-speed memory. The random access times to the main memory and the high-speed memory are 8.4 and 0.84 milliseconds, respectively. The machine has a photoelectric tape reader for input and a high-speed tape punch for output.

A standard digital-computer program was available for use in the solution of the trajectory equations given in the Appendix. This program was used in two ways in the comparison study. First, a standard trajectory was run, using an established integration interval.[5] (From previous studies, it had been determined that further reduction of integration intervals resulted in no significant improvement in accuracy.) Such a solution was then established as the standard of comparison for all other methods included in this study. Second, a trajectory was computed by the standard procedure, but the integration-step size was increased as much as possible while keeping the percent error in missile range under a specified amount. The results obtained from the digital computer using the standard program described in reference 5 were considered correct, and all deviations were based on these results.

Comparative Performances of Computers and Computer Programs

Solutions to the trajectory problem which served as a vehicle of comparison (see Appendix) were obtained by using (1) the standard digital-computer program, (2) the digital-computer program employing coarser integrating procedures, (3) the DEPI system, (4) the analog computer, and (5) the digital differential analyzer.

In the digital-computer solution with lowered accuracy requirements, the chief departure from the standard procedure was to allow integration-step sizes of large magnitude within the Runge-Kutta procedure. Replacement of the Runge-Kutta procedure by simpler methods of integration was not nearly so effective in reducing the computing time while retaining an accuracy corresponding to analog-computer solutions.

The standard digital-computer trajectory computation required 8 min (minutes); the coarse-integration procedure, 2 min; the DEPI solution, 4 min; and the analog solution, 0.5 min. The time required for obtaining a trajectory solution on a digital differential analyzer depends largely on the scaling of the independent variable. Three scalings were chosen so as to obtain solutions up to burnout in 3 min, 30 min, and 5 hr (hours). The DDA program failed to operate satisfactorily over the 5-hr time interval.

The digital computer and analog computer cost, respectively, about $125,000 and $200,000. The DDA costs about $100,000.

The results obtained in the various solutions are compared in Tables I and II and presented graphically in Figs. 1–4. Fig. 1 shows the velocity to burnout for

Table I. Comparison of Velocity Data

Time t, Sec	Missile Velocity V, Ft/Sec: Digital Computer (Standard Procedure)	Digital Computer (Coarse Integration)	DEPI	DDA 1	DDA 2	Analog Computer	Deviation from Standard Velocity, Ft/Sec: Digital Computer (Coarse integration)	DEPI	DDA 1	DDA 2	Analog Computer	Deviation from Standard Velocity, %: Digital Computer (Coarse Integration)	DEPI	DDA 1	DDA 2	Analog Computer
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	90	90	90	92	90	104	0	0	0	0	+14	0	0	0	0	+16.0
4	183	183	182	190	185	201	0	−1	+2	0	+18	0	−0.5	+1.1	0	+10.0
6	277	277	277	291	285	304	0	0	+6	+2	+27	0	0	+2.1	+0.7	+10.0
8	372	372	372	392	382	386	0	0	+14	+8	+14	0	0	+3.8	+2.2	+4.0
10	466	466	466	491	480	489	0	0	+20	+10	+23	0	0	+4.2	+2.1	+5.1
12	558	558	559	587	574	583	0	+1	+25	+14	+25	0	+0.1	+4.5	+2.5	+4.4
14	648	648	649	679	665	680	0	+1	+29	+16	+32	0	0.1	+4.5	+2.5	+4.9
16	734	734	736	767	752	750	0	+2	+31	+17	+26	0	0.3	+4.2	+2.3	+3.5
18	818	818	819	849	834	820	0	+1	+33	+18	+2	0	0.1	+4.0	+2.2	+0.2
20	898	898	900	927	912	900	0	+2	+31	+16	+2	0	0.2	+3.5	+1.8	+0.2
22	976	976	978	1000	985	990	0	+2	+28	+14	+14	0	0.2	+2.9	+1.4	+1.4
24	1052	1052	1054	1060	1055	1050	0	+2	+24	+9	−2	−0.3	0.2	+2.3	+0.9	−0.2
26	1098	1094	1101	1108	1098	1100	−4	+3	+40	+3	+2	−0.3	0.3	+3.6	+0.3	+0.2
28	1146	1142	1150	1136	1148	1155	−4	+4	+10	0	+9	−0.3	0.4	+0.9	0	+0.8
30	1208	1204	1213	1173	1214	1200	−4	+5	−10	+2	−8	−0.3	0.4	−0.9	+0.2	−0.7
32	1284	1280	1290	1222	1290	1295	−4	+6	−35	+5	+11	−0.3	0.3	−2.7	+0.4	+0.9
34	1374	1371	1381	1279	1377	1390	−3	+7	−62	+6	+16	−0.2	0.5	−4.5	+0.4	+1.2
36	1478	1475	1486	1343	1476	1470	−3	+8	−96	+3	−8	−0.2	0.5	−6.5	+0.2	−0.5
38	1597	1594	1607	1414	1588	1570	−3	+10	−135	−2	−27	−0.2	0.6	−8.5	−0.2	−1.7
40	1732	1729	1742	1493	1714	1710	−4	+10		−8	−22	−0.2	0.6		−0.5	−1.2
42	1883	1880	1894	1574	1851	1860	−3	+11		−18	−23	−0.2	0.6		−0.9	−1.2
44	2051	2048	2064	1662	1996	2045	−3	+13		−32	−6	−0.1	0.6		−1.6	−0.3
46	2238	2235	2252	1749	2155	2225	−3	+14		−55	−7	−0.1	0.6		−2.5	−0.3
48	2449	2445	2464	1837	2320	2420	−4	+15		−83	−29	−0.1	0.6		−3.4	−1.2
50	2684	2681	2702	1921	2494	2540	−3	+18		−129	−44	−0.1	0.6		−4.9	−1.6
52	2950	2946	2969	2000	2680	2950	−4	+19		−190	0	−0.1	0.6		−6.4	0
54	3248	3244	3269	2074	2879	3260	−4	+21		−270	+12	−0.1	0.6			+0.4
56	3585	3581	3609	2144	3082	2680	−4	+24			−5	−0.1	0.6			−0.1
58	3970	3964	3997	2203	3300	3950	−6	+27			−20	−0.1	0.7			−0.5
60	4410	4404	4441	2460	3525	4535	−6	+31		−885	+125	−0.1	0.7		−20.1	+2.8

Table II. Comparison of Range-Versus-Elevation Data

Digital Computer (Standard Procedure): Range x, Ft	Altitude y, Ft	Digital Computer (Coarse Integration): Range x, Ft	Altitude y, Ft	DEPI: Range x, Ft	Altitude y, Ft	DDA 1: Range x, Ft	Altitude y, Ft	DDA 2: Range x, Ft	Altitude y, Ft	Analog Computer: Range x, Ft	Altitude y, Ft
0	0	0	0	0	0	0	0	0	0	0	0
5	67	5	67	5	67		15		45	2000	5000
40	337	40	337	39	337	30	315	24	330	4000	10000
121	789	121	789	121	790	108	780	102	795	6000	14500
266	1422	266	1422	266	1423	258	1455	264	1440	8000	18000
487	2230	487	2230	488	2233	492	2310	486	2265	10000	20000
795	3207	795	3207	796	3212	798	3345	1816	3270	12000	23000
1199	4344	1199	4344	1202	4351	1212	4530	1230	4455	14000	25000
1708	5629	1708	5629	1712	5640	1722	5895	1752	5760	16000	28000
2327	7053	2327	7053	2332	7067	2328	7380	2382	7215	18000	30000
3062	8604	3062	8604	3069	8621	3054	8985	3120	8790	20000	32000
3918	10271	3918	10271	3927	10293	3888	10740	3990	10485	22000	34000
4900	12046	4900	12046	4911	12073	4842	12585	4968	12270	24000	36000
6000	13899	5998	13896	6014	13934	5916	14490	6054	14145	26000	38000
7202	15790	7197	15781	7220	15834	7050	16425	7248	16050	28000	39000
8524	17734	8515	17719	8547	17788	8280	18390	8562	17985	30000	41000
9985	19748	9973	19728	10014	19814	9618	20355	10020	20040	32000	43000
11607	21849	11593	21824	11643	21929	11076	22410	11634	22155	34000	45000
13413	24052	13396	24022	13547	24147	12654	24495	13428	24360	36000	46500
15427	26373	15407	26337	15480	26485	14388	26640	15426	26715	38000	48000
17673	28826	17651	28784	17737	28957	16290	28845	17628	29160	40000	50000
20179	31427	20154	31379	20254	31519	18348	31110	20082	31755	42000	51500
22973	34191	22946	34137	23060	34366	20598	33435	22800	34470	44000	53000
26086	37135	26057	37075	26188	37337	23028	35820	25782	37335	46000	54500
29554	40280	29523	40213	29672	40510	25650	38295	29088	40380	48000	56000
33419	43649	33386	43574	33555	43912	28452	40770	32706	43545	50000	57500
37728	47269	37691	47185	37884	47567	31458	43305	36648	46890	52000	59000
42533	51171	42493	51077	42712	51509	34632	45870	40962	50145	54000	60500
47896	55390	47853	55286	48102	55773	37998	48435	45648	54090	56000	62000
53890	59970	53844	59854	54126	60403	41508	50970	50742	57945	58000	64000
60603	64963	60552	64833	60874	65453	45306	53595	56172	61950	60000	65500
										62100	67000

the several runs; Figs. 2 and 3 give the deviation of the velocity and the percentage deviation in velocity, respectively. From these results, it is evident that the analog computer retains about 98.5 to 98% accuracy, that both DEPI and the coarsely integrated digital computer are still more accurate, and that the 30-min DDA run has a final accuracy of about 80%. The DDA solution is much more accurate when the rate of change of velocity is small, but compares very unfavorably when the rate of change of velocity becomes large. In the first 20 sec (seconds), where the rate of change

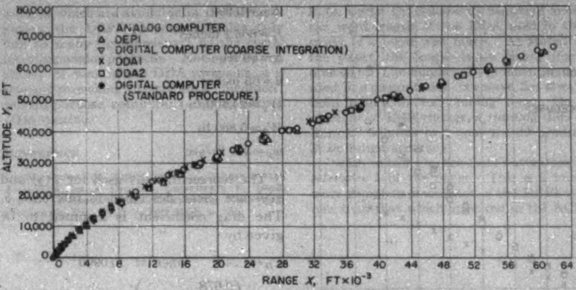


Fig. 1. Trajectory-problem solutions: velocity to burnout

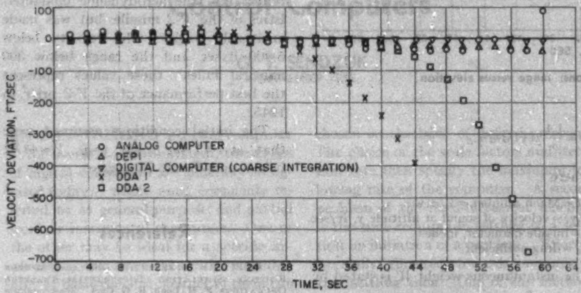


Fig. 2. Trajectory-problem solutions: deviation from standard velocity

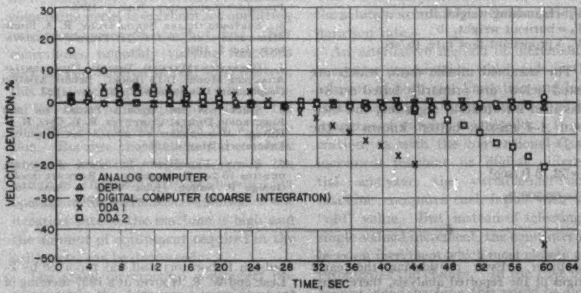


Fig. 3. Trajectory-problem solutions: percentage deviation from standard velocity

of velocity is small, the two DDA solutions do not coincide. Ideally, results obtained from a digital differential analyzer should be reproducible in the same manner as results from a digital computer. Certainly, some of the difference is due to the size of the integration steps taken in the two runs, but the greater part of the variation is probably due to the manner in which atmospheric pressure was simulated. This quantity was channeled into the computer via an optical curve follower. Past experience with missile-flight simulation indicates atmospheric pressure to be a critical quantity, and percentage variations in $p(y)$ appear as errors of the same order in the solution of the problem.

Fig. 4 shows a comparison of the range-versus-altitude curves obtained by each of the means used for solution of the problem. Although these results are considered critical from a missile-design standpoint, they are not critical with respect to the computation. The DDA results agree well with the other results until the disparity in velocity becomes evident at about 40 to 50 sec after launching.

Conclusions

When a high level of accuracy is not required in the solution of a differential equation, the analog computer is preferable to a digital computer because of its speed of solution and the facility with which parameters may be varied.

The DDA, which utilizes analog problem attack techniques, is at a disadvantage with respect to both the digital and analog computer because of the disproportionate amounts of time necessary to obtain analog computer accuracy in a problem of the type used as a vehicle of comparison. Newer types of digital differential analyzers would appear in a more favorable light due to the more refined integration schemes and parallel, rather than to the serial, problem solving capacity.

The DEPI system represents a means by which personnel not familiar with digital techniques may obtain solutions to differential equations. DEPI has not been used at the Jet Propulsion Laboratory because standard digital techniques are better.

Appendix. Missile-Trajectory Problem Used as Vehicle of Comparison

In the trajectory problem chosen as a vehicle of comparison for the computer study, the following equations of motion represent the flight of the missile:

$$\frac{dx}{dt} = V \cos \Theta$$

$$\frac{dy}{dt} = V \sin \Theta$$

$$\frac{dV}{dt} = g\left[\frac{F-D}{W} - \sin \Theta\right]$$

$$\frac{d\Theta}{dt} = +\frac{g}{V} \cos \Theta$$

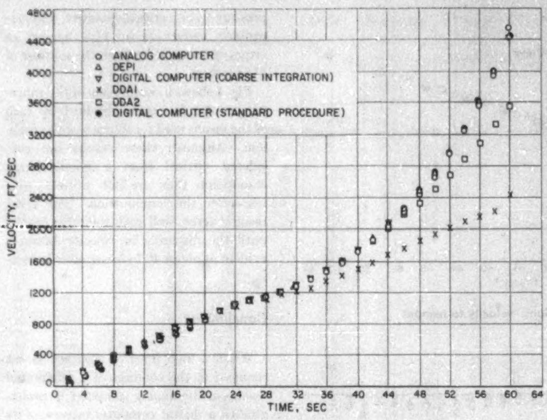


Fig. 4. Trajectory-problem solutions: range versus elevation

where

x = missile range, feet (ft)
y = missile altitude, ft
t = time, sec
V = velocity, ft/sec
Ⓗ = missile heading angle, degrees
g = acceleration due to gravity, ft/sec²
F = thrust, pounds (lb)
D = drag, lb
W = instantaneous weight, lb

The thrust F is stated by the following expression:

$$F = K_{jv}F_0 - f_e p(y) \quad \text{when } t<t_b$$
$$= 0 \quad \text{when } t>t_b$$

where

K_{jv} = percentage factor to account for jet-vane loss
F_c = vacuum thrust, lb
f_e = exit area, inches²
$p(y)$ = atmospheric pressure at altitude y, lb/inches²
t_b = burnout time, sec

The expression for drag D is

$$D = \frac{1.4}{2}\, p(y) M^2 d^2 C_D$$

where

M = Mach number = $V/a(y)$
$a(y)$ = velocity of sound at altitude y, ft/sec
d = missile diameter, inches
C_d = drag coefficient

The instantaneous weight W is stated by

$$W = W_g - w_p t \quad \text{when } t<t_b$$
$$= W_b \quad \text{when } t>t_b$$

where

W_g = launching weight, lb
W_b = burnout weight, lb
W_p = propellant flow rate, lb/sec

The assumed missile data, which are listed below, are primarily based on unclassified information regarding the German *A-4* missile, better known as the *V-2*.[6]

$g = 32.16$ ft/sec²

$F_0 = 68{,}500$ lb

$K_{jv} = 0.985$

$f_e = 665.1$ inches²

$t_b = 60$ sec

$d = 65$ inches

$W_g = 24{,}400$ lb

$W_b = 5{,}860$ lb

$w_p = 309$ lb/sec

The representations used for $p(y)$ and $a(y)$ are those described in reference 5. The drag coefficient is assumed to be given by

$$C_D = 1.246 \quad \text{when } M \;\; 0.0995$$
$$= 2\left(\frac{0.678}{M^2} + 0.2166\right) \quad \text{when } 0.995 < M < 5.0$$

This choice of the value for C_D does not reflect the true aerodynamic characteristics of the *V-2* missile, but was made arbitrarily to keep the velocity below 5,000 ft/sec and the range below 300 nautical miles; these values represent the best performance of the *V-2* prior to 1945.

The initial conditions assumed were that, at $t=1$ sec, $x=0$, $y=0$, $V=44.82$ ft/sec, and Ⓗ$=86.75$ degrees.

References

1. DEPI: AN INTERPRETIVE DIGITAL-COMPUTER ROUTINE SIMULATING DIFFERENTIAL-ANALYZER OPERATIONS, **F. H. Lesh, F. G. Curl.** *Memorandum no. 20–141*, Jet Propulsion Laboratory, Pasadena, Calif., Mar. 22, 1957.

2. ANALOG-COMPUTER FACILITY AT THE JET PROPULSION LABORATORY, **R. A. Bruns, J. J. Wedel, H. A. Brenneis.** *Report no. 20–89*, Jet Propulsion Laboratory, June 1, 1955.

3. SERVOMULTIPLIER ERROR STUDY, **R. A. Bruns.** *Memorandum no. 20–118*, Jet Propulsion Laboratory, Jan. 2, 1956.

4. OPERATION MANUAL, DIGITAL DIFFERENTIAL ANALYZER MODEL D-12 (book). **Bendix Aviation Corporation**, Los Angeles, Calif., Nov. 1955.

5. SIMULATION OF MISSILE FLIGHT ON THE ELECTRONIC DIGITAL COMPUTER, **R. E. Carr, H. B. Hilton.** *Memorandum no. 20–128*, Jet Propulsion Laboratory, May 1, 1956.

6. ROCKET PROPULSION ELEMENTS, AN INTRODUCTION TO THE ENGINEERING OF ROCKETS (book), **George P. Sutton.** John Wiley & Sons, Inc., New York, N. Y., first edition, 1949.

◆

Discussion

Ernest J. Schubert (Consultant, La Puente, Calif.): The pro's and con's of different computer concepts in real-time applications are often presented with emotional overtones. For several reasons, I lean more toward the arithmetic general-purpose computer for critical control applications, but feel the paper does not do the DDA justice.

From the section on "Comparative Performances" and Table I, it is apparent that the accumulated errors in the DDA become excessive and render both types DDA 1 and 2 inferior as compared with an analog computer. Even considering the early origin of the reported analysis, there must be some basic systems design mistake, possibly a too short word length. Can the author explain the reason for these poor results?

On several occasions this paper refers to the DEPI systems without giving any further details. In my opinion, the omission of this information, even as a reference, is regrettable because the DEPI system is one of the finest examples of operational simulation techniques. For the sake of completion, some of the highlights of this technique are quoted from the pertinent publication.[1]

The DEPI system is an interpreter written for the Datatron 204 at the Jet Propulsion Laboratory and was reported by F. Lesh and W. R. Hoover at a 1957 meeting of the Association for Computing Machinery (ACM). It simulates an analog computer using a fourth-order Runge-Kutta type of integration and comprises subroutines equivalent to square-root generators, resolvers, relays, dividers, function generators, etc. Each "component" representing one of the before mentioned functions has a 3-digit code and a corresponding memory location where its output is stored. Programming of DEPI consists of writing a sequence of control words corresponding to the analog block diagram. Because of the inherent capabilities of multiplication and division which are available in digital computers, the DEPI program contains only half as many "components" as the corresponding analog program. Noting the complexity of some of the DDA programs, it is thought-provoking to compare a "qualified" general-purpose computer program with a DDA routine.

REFERENCE

1. METHODS OF SIMULATING A DIFFERENTIAL ANALYZER ON A DIGITAL COMPUTER, F. Lesh. *Journal*, Association for Computing Machinery, Pittsburgh, Pa., vol. 5–3, July 1958, p. 281.

F. G. Curl: Dr. Schubert's comment anent the word length of the DDA used in this study is a point well taken. As stated in the paper, the length is six binary digits plus sign. The number of significant figures used in a computation, of course, has always been a function of running time. On a machine of 1949 vintage, a running time of from 30 minutes to 1 hour should produce 3- to 4-digit accuracy

I wish to call Dr. Schubert's attention to reference 1 of the paper. This is a much more complete description of the DEPI system than that which appeared in the ACM *Journal*. Although DEPI was developed at the Jet Propulsion Laboratory, it had certain serious limitations; chief among these was that it interpreted control words which represented analog connections. The DEPI system should have been written as a compiler rather than an interpreter, in that the control words could have represented differential equations rather than analog connections and the digital computer could then have connected the analog components in the most efficient manner. This was one of the principal reasons that DEPI was not at the Jet Propulsion Laboratory.

Results of Simulation Comparison of Control Computers

G. T. SENDZUK
NONMEMBER AIEE

IN CONSIDERING a digital computer for use in a control system, two types of digital computers are generally recognized today: whole word, commonly referred to as general-purpose, and partial word or incremental computers. One or the other may be ideal for a specific application, but, since they differ considerably in basic concept, neither of them can solve all problems.

The whole-number or general-purpose computer, in which no variables (with minor exceptions) are carried beyond one computing cycle, is excellent for operating on discontinuous or one-shot data. For continuous, smoothly varying functions such as encountered in real time control, this type of computation may be wasteful of time or equipment. Here, the incremental computer finds its application. Because the incremental machine concerns itself only with changes in the variables, and fewer digits are handled per operation, the incremental solution or the iteration rate of the machine is high and the amount of equipment required in the arithmetic can be decreased.

The usual incremental computers operate with the increment size fixed at values of plus or minus one and zero. However, because of the fixed increment size, the problem of rate limitation exists. Scale factors on increment size must be chosen on the basis of desired precision. The choice of the scale factors and iteration rate then specify the maximum following rate of the computer. A second problem of the fixed-increment machine is the time required in settling to a solution on initiation of a problem. Effort is normally directed towards minimizing this settling time. One of the methods employed is to preset initial conditions through the use of a supplementary whole-number computer or arithmetic section. The incremental computer then follows the solution accurately at the desired high iteration rate.

An alternative method of incremental computer organization which aids in the solution of both rate limitation and high speed slewing, is the employment of increments of varying magnitudes. In operation, as with the conventional fixed-increment machine or digital differential analyzer, the variable-increment machine compares each input with the "old" value. But, instead of selecting a single-valued increment, the computer selects an increment which most closely approximates the change. In the case of a binary coded machine, this would be analogous to selecting an integral power of radix 2 (±1, ±2, ±4, ±8, etc.). In some cases, the zero increment is also included.

In an attempt to aid in the process of selection for a given application a comparison of digital computers by simulation techniques was made by programming the operation of both incremental and whole-number machines on a general-purpose computer.

Control Computer Simulation

The simulation was performed using operating specifications of an existing whole-number machine and an existing variable-increment machine. Both machines were of approximately the same size and weight class (0.5–0.7 cubic foot and 30–40 pounds) employing equivalent clock rates and word lengths of 160 kc and 25 bits, respectively.

The whole-number machine was a 2-address machine employing a magnetic drum as its memory element. The variable-increment machine was a drum machine employing increments which varied in integral steps from $\pm 2^0$ to $\pm 2^{13}$.

Simulation of a fixed-increment machine was accomplished by assuming an an equivalent clock rate (160 kc) and word length (25 bits). The increment was fixed at $\pm 2^0$. The application chosen for simulation was one which, at present, is implemented by analog techniques. Digital techniques were considered to determine whether improved operation of the weapon system could be obtained.

The simulated problem is a weapon delivery problem involving the computation of a predicted aircraft weapon release

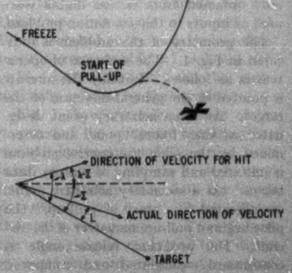


Fig. 1. Bomb geometry

Paper 61-724, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted April 4, 1961; made available for printing May 12, 1961.

G. T. SENDZUK is with the General Electric Company, Schenectady, N. Y.

$$T = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$b = -\left[V' - V_{EP} + W_R \cos(\lambda - \Sigma)\right]$$

$$a = \frac{1}{2}\, g \sin(\lambda - \Sigma)$$

$$c = D' \cos \lambda$$

$$\sin \lambda = \frac{\frac{1}{2}\, gT^2 \cos(\lambda - \Sigma) + \left[V_{EN} - W_R \sin(\lambda - \Sigma)\right] T}{D}$$

$$\sin(\lambda - \Sigma) = \sin \lambda \cos \Sigma - \sin \Sigma \cos \lambda$$

$$\cos(\lambda - \Sigma) = \cos \lambda \cos \Sigma + \sin \lambda \sin \Sigma$$

$$\cos \lambda = \sqrt{1 - \sin^2 \lambda}$$

$$\Sigma = \Sigma_0 + \int \frac{V \sin L}{D}\, dt$$

$$D = D_0 - \int V \cos L\, dt$$

LENGEND

T = TIME OF FLIGHT
V = AIRCRAFT VELOCITY
V' = CORRECTED AIRCRAFT VELOCITY
V_{EP}, V_{EN} = COMPONENTS OF EJECTION VELOCITY
W_R = WIND VELOCITY
D = RANGE TO TARGET
D' = CORRECTED RANGE
λ = PREDICTED RELEASE ANGLE
L = INDICATED RELEASE ANGLE

Fig. 2 (left). Bombing equations

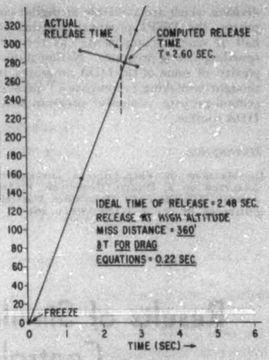


Fig. 3. High-altitude release of whole-number machine

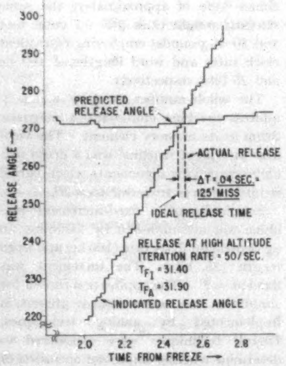


Fig. 4. High-altitude release of variable-increment machine

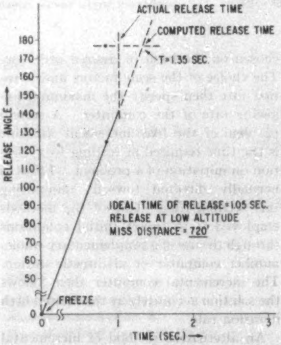


Fig. 5. Low-altitude release of whole-number machine

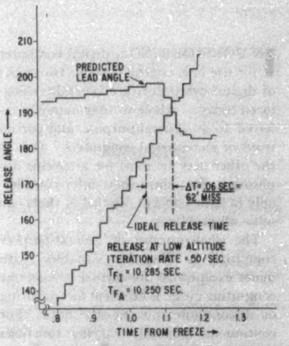


Fig. 6. Low-altitude release of variable-increment machine

angle and comparing it with the indicated release angle to affect a weapon release at the proper time. Recorded data in the form of aircraft velocity, altitude angles, etc., obtained from actual flights were used as inputs to the simulation problem.

The geometry of the problem is indicated in Fig. 1. The sequence of operation is as follows. Initially the aircraft is pointed in the general direction of the target. At some arbitrary point designated as the "freeze" point and determined by the pilot, the control problem is initiated and sampling of sensory data begins. At a secondary arbitrary point, as indicated by "start of pull-up," the pilot begins a pull-up maneuver of the aircraft. The predicted release angle is continuously computed and compared with the sample value of the actual or indicated angle supplied by the reference system.

Release of the weapon is accomplished at some time later when the conditions (predicted release angle equals indicated release angle) are met. This is shown in the lower part of the figure by the quantities L, for predicted angle, and λ, for indicated release angle.

Some of the equations used in the actual control problem are shown in Fig. 2. The parameters for which solutions are required are the values of L and λ. Because of the form of the equations, iterative procedures are best suited to obtain the desired values of L and λ. Definition of the equation parameters are given in Fig. 2. Actual problem length was fairly short, rarely running over several minutes.

The equations including appropriate ballistic and drag functions (not shown) were programmed on an International Business Machines Corporation (IBM) *650* general-purpose computer. The operation of the whole-number and incremental machines (both variable and fixed increment) were simulated using appropriate input data. Because of the shortness of the problem, the fixed-increment machine did not have sufficient time to settle and its outputs were not usable. The results obtained from the whole-number machine and variable-increment machine are explained subsequently.

Result of Whole-Word Versus Variable-Increment Simulation

The first part of the simulation was made using flight data for a high-altitude

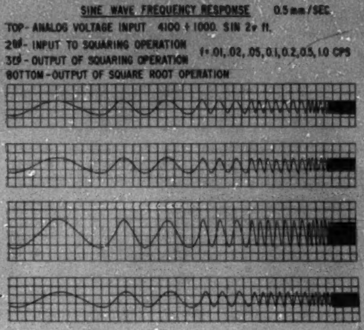


Fig. 7 (left). Sine wave response of variable-increment computer

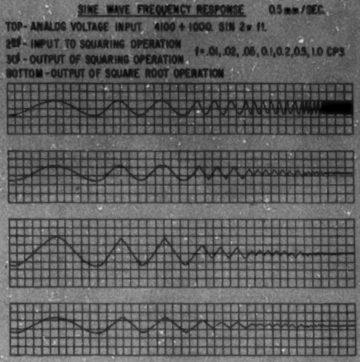


Fig. 8 (right). Sine wave response of fixed-increment computer

release. For the whole-number machine or general-purpose computer, the initial solution was obtained at approximately 1.5 seconds after problem initiation or freeze point. The second was obtained at approximately 2.9 seconds, too late to effect a release. The ideal point of release (calculated for the given input parameters) is 2.48 seconds; see Fig. 3. It can be seen that the error in time of release for the computer is approximately 0.1 second. For the simulated high-altitude release, this time error represents a miss distance of 360 feet.

The same high-altitude release parameters were used in the variable-increment computer operation simulation. The simulation general-purpose computer was programmed to function as an incremental computer with increments of variable magnitudes. The response of the variable-increment machine is shown in Fig. 4. After the initial slew at problem initiation or freeze point, the action of the computer in arriving at the release point is illustrated in Fig. 4. The iteration rate of the machine is 50 per second and incremental answers are generated every 1/50 of a second. Selection of increments of varying magnitude as the problem proceeds is shown by the varying height of the steps in the predicted and indicated release angle curves. The error in release time was 0.04 second. In a high-altitude release, this represented a miss distance on the ground of 125 feet.

It is evident that difference in miss distance for high-altitude release is relatively small for both computers. However, the differences in miss distance become significantly greater at lower altitudes as evident in Figs. 5 and 6. Low-altitude release input data were used in the whole-number simulation. The results are shown in Fig. 5. Note that the miss distance due to the time error in the release point is significantly greater.

The same problem operated on by the simulated variable-increment computer is shown in Fig. 6. Because of the ability of the computer to iterate at a high rate, the computer is able to arrive quite closely to the desired release point. The miss distance for the variable-increment machine is significantly less.

Additional Tests

The foregoing computer comparisons were made by simulation of computer operation of a general-purpose computer. An additional comparison, which may be of some interest, was made with the variable-increment machine and fixed-increment machine operation, using an existing variable-increment machine to perform the tests. The machine employs a 25-bit word and a 160-kc clock. It has an iterative rate of 130 per second.

In the fixed-increment mode, the machine uses increments of 0 and ±1. When switched to the variable-increment mode, increments of 0, ±1, ±2, ±4, ±8, ±16, ±32. . . . ±8,192 can be automatically selected by the machine.

Fig. 7 shows recordings made of the response of the computer operating in a variable-increment mode to a sine wave input of $4{,}100+1{,}000 \sin 2\pi$ feet, varying in frequency from 0.01 cps (cycle per second) to 1.0 cps. Trace 1 shows the analog voltage input. Trace 2 shows the analog signal converted to a digital number being used as the input to a squaring operation programmed in the com-

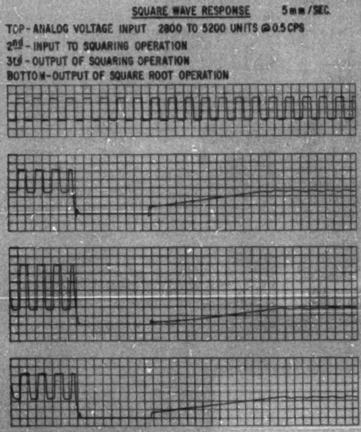


Fig. 9. Square wave response of variable-increment and fixed-increment computer

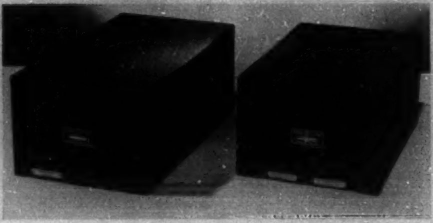


Fig. 10. The variable-increment computer

puter. The third trace represents the computed output of the squaring operation. Trace 4 represents the output of a square root operation using the output of the squaring operation as the input. Note the fidelity between the top trace showing the input signal and the output of the square-root operation of trace 4. Both are identical except for a transport lag introduced by the computer.

The same inputs were used with the computer operating as a fixed-increment machine using unity increments. Again the recorded traces represent the same quantities in the previous Fig. 8. Note that at approximately 0.02 cycle, there is already some evidence of rate limitation and at 0.5 cycle the outputs are practically nonexistent. An additional example of a comparison made between a variable-increment computer and fixed-increment computer is shown in Fig. 9. The example shown graphically illustrates the slew capability which can be built into an incremental machine with adjustable increment magnitudes.

The top trace of Fig. 9 is the analog voltage input in the form of a square wave. The step input varies from an initial value of 2,800 units to 5,200 units at a rate of 0.5 cps. Trace 2 shows the first three cycles of the input to a squaring operation with the machine operating in a variable-increment mode. Similarly, in traces 3 and 4 the first three cycles of the squaring and square root operations are shown. After the first three cycles, the values in the computer were reset and its operation switched to a fixed-increment mode. At about a third of the way through the recording, the initial conditions are re-established and the computer is then allowed to settle to a solution. Settling usually occurs at about the three-quarter point on the traces. Note the poor fidelity of the computer output (trace 4) with respect to the input.

The preceding recordings were obtained from specific algorithms which were part of a fairly large control problem programmed on the variable-increment computer developed by the General Electric Company. The actual computer exclusive of input-output equipment used is shown in Fig. 10. The computer employs a magnetic drum memory which revolves at 8,000 rpm giving it a iteration rate of 130 per second. The variable-increment computer employs increments which are selected automatically by the computer and can vary from 0 to $\pm 2^{13}$ in steps of integral powers of radix 2, i.e., $\pm 2^0$, $\pm 2^1$, $\pm 2^2$, $\pm 2^3 \ldots$, etc. By employing variable-increment magnitudes, the computer is capable of following at 130 units per second or in the slew condition, at better than one million units per second. This slew capability enables the computer to perform logical operations quite readily, thus making it usable for many applications where decision operations are required.

Conclusions

The results of the simulation are not all conclusive. For the problem simulated, the variable-increment machine performed better than a fixed-increment or DDA machine because of its ability to slew and settle more rapidly. The variable-increment machine also showed slight edge over the whole-number machine. But the difference was small. The decision as to which type of computer should be used in a given system is difficult to determine and may be dependent upon many other factors, such as ease of programming, mode switching, and decisions to be performed.

◆

Discussion

G. L. Hollander (Fullerton, Calif.): The work of the author has two facets of broad interest: (1) the actual method of the simulation and (2) general conclusions concerning the efficacy of whole-number and incremental computers for airplanes. A brief description of his simulation method, or reference to a specific publication, would not only satisfy the first facet but would also allow the reader to judge the validity of the author's general conclusions. The results depend heavily on how cleverly each simulated computer was programmed.

Dissimilar machines are best compared on their performance per unit cost, where "cost" includes drawbacks of size, weight, etc. The paper implies equal cost for the machines by mentioning equivalence of size, weight, clock rate, and word length. This approach is valid if all factors are equivalent. However, the reader would have to draw different conclusions if both machines did not received equal design effort or were based on different stages of technology. The variable-increment computer pictured in Fig. 10 of the paper had been described as a semiconductor machine.[1] Did the whole-number machine also use semiconductors or was it an earlier computer? Minor changes in technology can easily double the entire cost.[2]

Incremental and whole-number computers cannot be compared on the basis of a single routine. Any realistic application involves several tasks,[3] some of which are more suitable for the incremental, others for the whole-number computer. Furthermore, other factors must be considered. The author might like to contrast his results with those of Dickinson[4] and Shackell and Tryon,[5] particularly with the latter's estimate that the variable-increment computer is almost twice as unreliable as the other two types.[5]

The subject of the paper is significant. Further elaboration can make the information valuable for future work.

REFERENCES

1. A VARIABLE INCREMENT COMPUTER, G. T. Sendzuk. "Computers in Control," *AIEE Special Publication S-132*, Sept. 1961, pp. 112–20.

2. *Closure* of A COMPARISON OF DIGITAL DIFFERENTIAL ANALYZER AND GENERAL PURPOSE EQUIPMENT IN GUIDANCE SYSTEMS, M. M. Dickinson. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 80, May 1961, p. 192.

3. REAL TIME AIRBORNE DIGITAL COMPUTERS, Gerhard L. Hollander. *ISA Journal*, Pittsburgh, Pa., vol. 5, no. 12, Dec. 1958, pp. 26–32.

4. A COMPARISON OF DIGITAL DIFFERENTIAL ANALYZER AND GENERAL PURPOSE EQUIPMENT IN GUIDANCE SYSTEMS, M. M. Dickinson. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 79, 1960 (Jan. 1961 section), pp 706–08.

5. THE RELATIVE MERITS OF INCREMENTAL AND CONVENTIONAL DIGITAL COMPUTERS IN AIR-BORNE REAL-TIME CONTROL, S. M. Shackell, J. G. Tryon. *Ibid.*, Sept. 1960, pp. 393–400.

Educating Electrical Engineers for Professional Careers

SIMPSON LINKE
MEMBER AIEE

IN ANY current discussion of engineering education, particularly electrical engineering education, inevitably a conflict arises between those who wish to inject more science and mathematics into curricula and others who insist that engineering colleges should train engineers, not scientists. This paper supports neither side of this controversy but presents the arguments for the following thesis:

The present emphasis on science and mathematics in electrical engineering curricula is creating a research-oriented type of undergraduate education that may lead to inadequate preparation for professional careers. Since 80 to 90% of the undergraduates intend to follow professional rather than research or teaching careers, re-evaluation of the goals of undergraduate education is required.

Elements of the Problem

The charge of inadequate education cannot be based upon the proliferation of science and mathematics in electrical engineering curricula. There can be no doubt about the necessity for increased emphasis on these basic subjects in this age. College courses should concentrate on broad functions that are applicable to all branches of electrical engineering and devote less time to specialized devices, design, shop work, and so-called technical drudgery. Unfortunately, as a result of this scientific attitude, engineering courses are stressing mathematical analysis, based on idealized conditions and linearized parameters, rather than emphasizing the influence of physical limitations. Thus, the student assumes that everything works exactly according to theory. When these conditions are coupled with the prevalent philosophy, that engineers are only required to visualize and expound problems and that the actual final solutions are trivial, it is possible for many nonresearch-minded graduates to become mathematical robots, facile with differential equations, but completely out of touch with the physical realities. Such men will be neither scientists nor engineers. They know that they are inherently without scientific creativity and their education has indicated that engineering creativity is somehow beneath them.

The problem of the education of the professional engineer becomes even more complex when the characteristics of the present-day engineering student are considered. Many of these students are products of progressive education that has failed to tax their mental capabilities. These young men have had their early education "spoon-fed." They have been allowed to do only those things that they like to do and to avoid unpleasant, routine, and distasteful tasks. Many have decided to study engineering because of their natural aptitude for mathematics and physics. Others may have been influenced by the lure of high wages. Undoubtedly these young men are not very different from their predecessors of a decade or two before. Their attitude, however, differs greatly. They complain about being required to do unpleasant things. Courses in machine shop, surveying, drafting, fluid mechanics, and heat power may seem like a waste of time, particularly if they have heard faculty members share their view. Unfortunately, this attitude extends to such courses as elementary circuit theory, where it is often necessary to use numbers, utilize the slide rule, and obtain answers that must be interpreted with some engineering judgment. Enlightened instructors have abandoned the technique of the rote exercise and the dull repetition of routine problems. However, a certain amount of numerical work still remains in courses that retain their engineering flavor. One professor reports the amazement which greeted his requirement to his computing classes that they solve a reasonable circuit problem, involving complex algebra, down to the last detail. "Do you want us to work it all the way out?" he was asked. While the boys labored, he could hear them muttering to each other that, "This dog-work is for peasants."

In one elementary machinery course two students incorrectly wired a d-c motor-starter set under the supervision of a novice instructor. When the resulting short circuit kept blowing the circuit breaker and the instructor could not immediately spot the trouble, the two future engineers complained that they were wasting their time on bad equipment and left the laboratory in a huff.

A senior-project student, working on a computing device, obtained a rather large and complex commercial electronic power supply that could be applied to his project with some modifications. Although complete wiring diagrams were available, the student insisted that he would require the services of a technician to trace the circuit and find out which relays were blocked. The boy's computer circuitry was much more involved than that of the power supply and a few hours of effort would have removed all the difficulties. However, he strongly indicated that he was above tearing into the unknown circuit and making it work.

These incidents, though isolated, are indicative of the general attitude of today's engineering undergraduate. Frequently, this attitude is encouraged by young, sometimes brilliant, faculty members who are recent graduates and who have never engaged in practical engineering work. These young professors often are doing research work of their own, sometimes under generous government contracts, where factors of economy, reliability, and engineering judgment may be secondary. Their efforts usually are in areas where idealized and linearized parameters may be assumed with confidence. They rarely encounter such troublesome materials as silicon steel in their research work. These men are engineering scientists for whom the differential equation and the digital and analog computers are invaluable tools. Inevitably, when they instruct undergraduates their personal attitudes are communicated. However, only the upper 10% of each class can be expected to follow in their footsteps. The rest may be stimulated but will undoubtedly obtain a distorted view of engineering work. It must appear to the remaining 90% that their profession offers either a life of technical drudgery or stimulating discovery. If they find that they are unable to maintain the high standards required in research and graduate work, they may be condemned to the purgatory of ordinary engineering. Their lack of motivation for achieving technical excellence, engendered by the frustrations of

---

Paper 61-733, recommended by the AIEE Education Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 7, 1960; made available for printing April 25, 1961.

SIMPSON LINKE is with Cornell University, Ithaca, N. Y.

their engineering education, has led many of them to decide to become sales engineers, application engineers, engineering managers, engineering administrators; anything but professional engineers.

Criteria for Engineering Education

The problems of evaluating engineering education are complex; it took 18 months to draft the Grinter Report for the American Society for Engineering Education. Every electrical engineering faculty in the country is endeavoring to find answers to such questions as optimum number of semesters, ideal curriculum content, importance of the humanities, and related problems. It would, therefore, be foolhardy to suggest pat solutions to the multiple problems of undergraduate education. The problems vary with each institution and their severity depends on the temper of the faculty and student body. However, certain published commentaries on engineering education may be of use in providing guideposts for action.

In a recent editorial,[1] Dr. H. G. Booker, Director of the School of Electrical Engineering at Cornell University, made the following two important statements:

> 1. The principal function of undergraduate education should be to develop the mind of a student by having him think through things that other people have frequently thought through before.
>
> 2. There is nothing wrong, or even unduly specialized, about engineering as a basis of education, undergraduate or graduate. Engineering provides a wealth of material for developing the minds of students on the basis of both well-known and novel ideas.

These statements might be termed the "Booker Criteria for Engineering Education." Although general in nature, they serve as an appropriate yardstick for the evaluation of any facet of the educational process.

Application of the criteria to the problems of educating electrical engineers for professional careers removes the confusion that has clouded certain issues. For example, there has been considerable agitation to banish the conventional machine laboratory on the grounds that the devices are obsolete and that their development has reached a plateau. There can be no argument with such a proposal if machinery is still to be studied amid a mass of winding detail and specialized design techniques for the purpose of compiling voluminous data in order to draw dozens of characteristic curves. Such practice violates criterion 1 for rarely does such detailed work develop the student's mind and certainly it is not conducive to making him think, except in a mechanical way. On the other hand, if the machine is studied as an actual physical device from the standpoint of the relative interaction of the various components and parameters and if the student is required to think through the fundamentals as Faraday, Tesla, and Lamme did, then the study of the particular device is educational. Criterion 2 is also satisfied through the broadening influence of the study of the historical development of the machine and by allowing the student to come into contact with the reality of an engineering device through observation of the manifestations due to nonlinearities and the interpretation of deviations from theoretical predictions. If he is disturbed by the inaccuracies of his measurement devices, intrigued by the difficulties of analysis of ferromagnetic circuits in motion, and appalled by his lack of knowledge of magnetic materials, he is receiving engineering education of the purest kind. The main point is that his mind is being sharpened on old ideas so that at some future time he can contribute new ideas of his own.

The process of idealizing and linearizing the machine so that it may be analyzed by abstract mathematics is considered good educational practice only after the student is sufficiently sophisticated to comprehend the implications of his simplifying assumptions. To limit machine theory to the mathematical approach is poor educational practice. Valuable thought processes are bypassed and the important bridge from past to present thought is never crossed, but perhaps worst of all, the student is frequently left without a clear understanding of the limitations of his knowledge. The more discerning students exhibit a feeling of disappointment at this treatment. "I applied the equations of motion," said one graduate of this technique recently, "but I didn't learn much about machinery."

Granted that the object of electrical engineering education is not necessarily to learn about machinery, nevertheless a curriculum that fails to provide the student with a firm physical foundation upon which to build his mathematical concepts is comparable to a medical education that teaches the technique of surgery without anatomy. Perhaps the would-be surgeon learns some anatomy in the process but no one would care to find himself on his operating table.

Needless to say, these comments are applicable to other branches of electrical engineering undergraduate education. For example, a study of the characteristics of a vacuum-tube amplifier without an understanding of the fundamentals of physical electronics cannot be justified under either criterion. In short, engineering education based principally on mathematical analysis is unable to produce graduates who can apply their knowledge with judgment and, as the Engineers' Council for Professional Development defines an engineer, who can "develop ways to utilize, economically, the materials and forces of nature for the progressive well-being of mankind."

The Laboratory Stem

Present engineering education literature abounds with statements about the core curriculum. In contrast, relatively little attention is given to the laboratory. This would indicate that current trends are toward reduction in laboratory time with limited study of instrumentation and adoption of a policy of student *laissez-faire* in the choice of undergraduate laboratory projects. The already-mentioned reduction of power machinery to bench-scale devices and a general attitude of condoned laboratory carelessness, based on the belief that the liberated, uninhibited student will learn more rapidly by making mistakes, is also advocated.

Perhaps these attitudes have developed because of the general feeling that practical or specialized training should be given to the graduate in the cadet programs of industry and that only mathematics, physics, and engineering fundamentals need be taught in college. The fallacy in this is that industry does not care about developing the mind of the student. It is assumed that this has already been done by the school. Training in the specialty, not education of the individual, is what the cadet master has in mind. If the college has failed to instill the necessary engineering disciplines with standards of accuracy and comprehension of physical and material limitations plus the desire for technical excellence in the student, it is too late. In isolated instances an individual may possibly achieve these goals in industry. But how much broader would his education be if he had been exposed to these concepts in a stimulating and critical academic atmosphere for four or five years.

To provide adequate education for all graduates in electrical engineering, every core curriculum should have grafted onto it a sturdy laboratory stem. This program should parallel and complement the theoretical work throughout the required portion of the undergraduate program. A list of the desirable features of such a program follows:

1. Routine experiments, rote taking of data, repetitious computation, and standardized techniques should be avoided.

2. Theory and application of electrical measurements should develop as the student progresses.

3. The student should have a thorough understanding of the theory, behavior, and limitations of each instrument. There should be no "black boxes" in the laboratory.

4. Introduction of sophisticated devices into measurement should be delayed until the student can fully comprehend them.

5. Analysis and measurement of the characteristics of the various ferromagnetic materials and devices should be included.

6. A thorough investigation of the electrical behavior of real devices, not idealized analogs, should be undertaken.

7. The development of engineering creativity should be emphasized.

8. Provision for students should be made so they may develop ideas with minimum assistance from instructors. However, free selection of topics for study should be avoided until the senior year.

9. Firm laboratory rules should be established. Proper instrument care and personal safety should be stressed.

10. Orderly, well-organized techniques should be encouraged. A blown fuse should be considered the result of unwarranted carelessness until proved otherwise.

Throughout these laboratory sessions the emphasis must be on complete comprehension of all aspects of each project. The experiments should be arranged so as to place natural obstacles in the path of the students so that they will be forced to exercise their growing engineering judgment in order to overcome them. There should be ample opportunity for the observation of phenomena that deviate from the predictions of theory. Mathematical analysis should be encouraged whenever possible. When mathematics cannot be used, physical analysis should be employed.

Examinations on laboratory work should be creative. Problems that depend on mathematical derivations should not be used. The problems should be engineering problems that require thought, analysis, understanding of concepts, and, of course, good judgment. Problems that require the mere parroting of material absorbed in class should always be avoided.

It is unlikely that the students will react favorably to such a program at first. Their backgrounds and mental attitudes will undoubtedly cause them to be unreceptive to such treatment. If a parallel theoretical program is equally challenging, though, there will be a number of important benefits. The incompetent students and the would-be managers and salesmen will be eliminated early. The serious students will have ample opportunity to decide whether they wish to follow research or technical careers. By proper choice of electives, the students in each group will be able to prepare themselves for the careers of their choice. In either case, their undergraduate education will have been thorough and will have given them adequate preparation for the future.

Reference

1. What Is Wrong with Engineering Education, H. G. Booker. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 42, no. 3, Mar. 1954.

◆

Discussion

Eric T. B. Gross (Illinois Institute of Technology, Chicago, Ill.): Professor Linke has presented a paper dealing with important questions. Strange things are happening at many colleges where "new" curricula are developed, all looking fine on paper; however they cannot be carried through as intended because of the limitations of students and faculty. Unfortunately, what Professor Linke has stated is only too true, and he is very courageous to tell his story. But engineering education should certainly have stressed mathematical analysis and emphasized physics and chemistry to a much greater extent a long time ago. Technical universities abroad have been enabled to do this since their establishment as their students come from special preparatory schools which bring a limited number of students to an educational level that corresponds to at least the end of our sophomore year. We could surely do the same with freshman on similarly advanced levels if our high school produced such freshmen (though lately some have accelerated programs on a somewhat advanced level). Consequently, there is insufficient time in the colleges for a thorough job, and emphasis has been shifted from the "why" to the "how." Students can be taught mathematical manipulations of an "advanced" nature, but not much else. It has been observed that most first year graduate students have no trouble in solving a simple R-L circuit by operational methods, but many have overwhelming difficulty solving the same problem by elementary calculus. In a recent lecture someone pointed out that current graduates compare very unfavorably with those of earlier periods in professional engineering work, but were superior in research and development assignments. At first this seemed quite satisfactory, but a further analysis showed that all these research and development boys added at least one year of postgraduate university education to the B.S. level. They were not only the better students, but also those whose formal studies had continued on to advanced levels. Of course, they were qualified for research and development work. All this fits into Professor Linke's presentation. Is it really the purpose of our colleges to educate only the top 10% and to teach the majority only the "how" of mathematical manipulations?

Philip Sporn suggested in a recent address[1] that there have never been too many engineers who had become college professors, but that the latest developments are producing still fewer engineers on college faculties. As soon as this situation changes, there will no longer be a need for Linke's suggested re-evaluation; significant changes will certainly lead to newer curricula, based on new "revolutions" in education.

For some reason, Professor Linke has paid particular attention to electric machines (electromechanical energy conversion devices, if you please). I am sure that those university professors who are able to engender enthusiasm in their students during a machinery course never bothered with such details as winding designs or the various methods of finding the value of regulation. What should be covered in such courses has recently been summarized by B. Adkins[2] and need not be repeated here. Adkins has clearly indicated what can be properly handled in an undergraduate course, and what post-graduate education should cover.

Mathematical analysis alone will never suffice in engineering. It must either be accompanied by or followed by physical analysis so that the results are meaningful.

References

1. Science, Technology, and the Training of Engineers for the World of Tomorrow, Philip Sporn. *Electrical Engineering*, vol. 80, June 1961, pp. 409–13.

2. The Rotating Electrical Machine in an Electrical Engineering Degree Course, B. Adkins. *Ibid.*, May 1961, pp. 347–49.

W. R. Le Page (Syracuse University, Syracuse, N. Y.): Professor Linke illuminates the alarming trends in modern engineering education, indicating how easy it is for a student to graduate without developing a realistic appreciation of the relative position of theory and experiment.

It was pointed out in the paper that theory applies to idealized situations, but Professor Linke does not stress the fact that all theory utilized in physical science (other than pure mathematics) is based on experiment. Thus, experimentation is always necessary to establish the theoretical models, and to provide knowledge as to their accuracy and limitations. A student should be sensitive to this fact to a greater degree than is possible if he is only taught it verbally. Furthermore, except in a few rare cases, theory is not adequate to attain a complete solution to a design problem. Both of these points give weight to the argument that not only should the educational environment provide an adequate amount of work in the laboratory, but this work should be properly related to theory. This relationship is one of interdependence, each supporting the other, rather than the often misconstrued attitude that experiment merely illustrates theory. The desirability of maintaining a close integration between theory and experiment argues against an overconcern about teaching a given topic in a theory course prior to its use in the laboratory. Laboratory experimentation should always be done in the light of the best theoretical knowledge, and the attitude in the laboratory should be one of seeking appropriate theoretical knowledge at the time it is needed. If this attitude is persistently maintained, we should graduate engineers who are competent to bring the most appropriate methods to bear on the solution of practical problems.

Lincoln R. Samelson (Editor, *Insulation*, Libertyville, Ill.): Professor Linke made an excellent point when he stated that a re-evaluation of the goals of undergraduate electrial engineering education is needed to prevent 90% of the students from entering a "never-never" land of theory with the danger of becoming mathematical robots. He stated that the present emphasis on science and mathematics in electrical engineering curricula is creating a research-oriented type of undergraduate education that may lead to inadequate preparation for professional careers in engineering.

As pointed out by Professor Linke, such education may make it possible for many nonresearch-minded graduates to become facile with differential equations but completely out of touch with the physical realities of engineering problems. I agree wholeheartedly with the thoughts of Professor Linke. In recent years the pendulum has swung too far toward science for *all* students; an educational compromise is necessary for the many students who plan to be engineers rather than scientists. Otherwise, a considerable number of these graduates might turn out to be poor engineers or poor scientists or even turn to other professions.

The suggestions of Professor Linke do not have to mean any lessening of the training of scientific manpower; they merely mean that we should more realistically evaluate the purposes of our undergraduate educational programs in order to educate better engineers while still providing a firm foundation for those who wish to pursue graduate studies leading to research and scientific work.

Simpson Linke: I wish to thank the discussers for their interest and for their kind comments. The greatest value of papers of this type lies in the discussions that they evoke. It is noted with regret that some of the opponents of the viewpoints expressed in the paper have not contributed their rebuttals. Perhaps these opposing views will be presented soon in some other format. Certainly the author hopes that debate on this important engineering education topic will not end with the publishing of this paper.

Professor Gross suggests that an influx of practicing engineers into engineering faculties will eliminate the need for the re-evaluation of engineering education called for in the paper. Perhaps significant improvement can be achieved in this way if people from industry are brought in as Visiting Professors for periods for one year. Some engineering colleges have followed this practice in a limited way with decided mutual benefit. Conversely, a sabbatical year spent in industry by an engineering professor is an enlightening experience. Long-term exchange programs of this type would certainly help to keep engineering curricula in touch with reality.

Professor Gross also comments on the stress given in the paper to the study of electric machinery. It was not intended that the statements should imply that machinery is of greater or lesser importance than any other pertinent subject in an electrical engineering curriculum. The important thing is that, whatever the subject, the laboratory work should require the student to engage in the creative thought necessary to overcome the natural obstacles that are frequently encountered in all phases of engineering work.

It is on the latter point that Professor LePage's comment is particularly pertinent. Engineering colleges are derelict in their duty if their graduates have not gained the benefit of laboratory experimentation that illuminates the theory of physical science. However, this kind of laboratory work requires a degree of sophistication that can only be obtained by practice and by attention to the details and limitations of instrumentation. It is for this reason that the list of suggested laboratory procedures is included in the paper. While early laboratory work can and should include elements of true experimentation, it is also necessary to teach disciplines that will ensure meaningful results in the students' later work. Possibly much of the experimental laboratory work should be relegated to the senior year. The Senior Projects, prevalent in many schools, is an important example of this philosophy.

The author is grateful to Mr. Samelson for his endorsement of the views expressed in the paper. Continued interchange of thoughts and ideas between educators aud representatives of industry will undoubtedly reveal areas of agreement and disagreement about engineering education that should eventually result in significant improvements in curricula.

Motivation Through Challenge

W. C. JOHNSON
FELLOW AIEE

P. R. CLEMENT
NONMEMBER AIEE

EVERY TEACHER IS well aware of the importance of motivation in the performance of his students. It is a matter of common observation that a student of superior ability who loses his desire to learn quickly becomes an academic liability. Achievement seems to be something like the product of natural ability times motivation, integrated over the period of learning. Although the source of motivation is sometimes beyond the instructor's control, on the average it is the teacher himself who must take the responsibility for generating in his students not only the desire to learn but the will to do the work of learning. The thesis of this paper is that a strong intellectual challenge is the most effective means of motivation, and that electrical-engineering education over the past few years has developed in precisely this direction. One such program, based on a new kind of introductory course, is described.

An outstanding characteristic of the young man is his willingness to accept a challenge and his desire to test himself on difficult tasks. In fact, if challenges do not naturally arise, he is likely to manufacture them. He has the belief, often justified, that he has the ability to accomplish great things in the future. Furthermore, he has the important advantage over his elders of having the time to prepare himself for distant goals.

Most teachers will agree that another outstanding characteristic of the college student is the basic desire to learn and to fit his knowledge into a coherent whole. The student of college age can assimilate knowledge and master ideas at the greatest rate of his life. His intelligence is fully developed, his mind is uncluttered, and his approach to new ideas is fresh and uninhibited.

People, it is well known, are creatures of habit. Good habits of intellectual discipline, like muscles, are developed more easily in young people than in old, and this development is obviously one of the important objectives in a college education. Intellectual skill, fully as much as physical skill, is acquired only through practice and exercise.

Paper 61-735, recommended by the AIEE Education Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted February 13, 1959; made available for printing June 14, 1961.

W. C. Johnson and P. R. Clement are both with Princeton University, Princeton, N. J.

So far we have only been calling attention to, and trying to formulate precisely, what every teacher knows but what all of us are sometimes inclined to forget: that young people require challenges, that they are the most competent learners, and that it is an important function of the teacher to provide the challenge. It is not so likely that the teacher will neglect to enforce intellectual discipline, for this is his traditional job. But each of us, being of the older generation, is likely to discount the competence of youth and thereby fail to offer the stimulation and challenge that such competence deserves, and fail to require the full measure of intellectual discipline of which the young person is capable.

The developments that have taken place in electrical-engineering education over the past few years are in the direction of greater challenge to the student, and it is the effect of these new programs and courses on the student which confirms the belief in challenge as a motivating factor. The general characteristics of the new programs have been an increased emphasis on the scientific aspects of electrical engineering and on the introduction of sophisticated ideas at an early stage in reasonably rapid order. There are many who, from experience, will agree to the stimulation afforded to the student through his grasp of ideas that were formerly thought to be the province of the graduate student, or indeed were not taught at all. The main problem is not the ability of the student to assimilate ideas; the problem is to formulate the ideas clearly and correctly without overshooting the mathematical preparation of the student. If all high schools taught good courses in the calculus, the flow of ideas could be much faster in the first two years of college.

An introductory course in electrical engineering must wait for the student to get reasonably well along in the calculus and so is generally given, or started, in the sophomore year. There are at least two schools of thought concerning the first course. One holds that circuit theory is the logical starting point, both because of its importance in the field of electrical engineering and because its mathematical requirements and its order of difficulty seem appropriate to the student at this point. Another school of thought believes that the basic concepts of electricity and magnetism form a more appropriate starting point because they are more fundamental than circuit theory. They are introduced in a physics course often given in the freshman year and can be carried forward immediately in the sophomore year; the ideas, if formulated properly, are appropriate to this level of learning. The physical ideas of circuit theory are then established preparatory to the exploration of the mathematical aspects of networks.

Each of the foregoing introductions, the physical aspects of field and circuit theory and the mathematical aspects of network analysis, can provide challenge for the student. The authors believe in starting with the field approach, and will conclude this paper with a brief description of a 5-year experiment made from this point of view.

As is true in many schools, our freshmen take elementary differential and integral calculus with analytic geometry and general physics with emphasis toward the end on electricity and magnetism. When he begins the sophomore year the student knows what a derivative and an integral are, and he can differentiate and integrate simple functions, but his mathematical background has not reached the point where he can make use of partial differentiation and the other mathematical apparatus often associated with a sophisticated approach to electromagnetic theory. The authors feel, however, that this can be a blessing in disguise, since the ideas of electromagnetic theory are physical, and too much mathematics at the beginning can obscure the physical concepts. The aim is to present the physical ideas correctly and in such a form that the student will not need to unlearn them later. In such a presentation the student can see that mathematics of a more advanced kind would permit him to manipulate the physical ideas more adroitly and to solve problems of greater complexity, thus motivating him in his study of mathematics. The ideas of differential and integral calculus are used freely and the multiplication of vectors is introduced, but vector analysis is considered no further than this. We are therefore limited to the integral formulation of Maxwell's equations and to examples and problems which have a high degree of spacial symmetry. Fortunately, many interesting problems can be solved within this limitation. Since the student's mathematics is used soon after he learns it, one must keep informed about what the mathematics department is doing so that one will not assign problems that require, for example, previously unexplained techniques of integration.

The material on electricity and magnetism occupies about half a semester and includes as examples and problems many applications in engineering. Immediately thereafter the basic principles are applied to two areas: to the conversion of energy with emphasis on electromechanical transducers, and to the theory of lumped circuits, with particular attention to the relation between the physical concepts and their mathematical formulation. Thus, the treatment proceeds from field ideas to circuits and physical apparatus, and to their mathematical models. With the circuit relations formulated, attention turns to the analysis of networks, starting with network topology and extending through zero-pole ideas. The material of the 1-year introductory course stops just short of the Laplace transform.

In the accompanying weekly laboratory periods, experiments are performed only in alternate weeks in the first semester to allow the class work to get ahead, and are held somewhat more often during the second semester. The remaining laboratory periods are used for problem solving, particularly for problems which are more difficult mathematically or in which the student requires some guidance in the exercise of judgment, as in making a design. The laboratory proper stays as nearly as possible in step with the theory taught in class. In both problems and laboratory, the objective is to show the practical use of the theory and the limitations of the models which one uses to analyze physical situations. In one experiment the student uses a simplified oscilloscope to measure the forces exerted on a stream of electrons by electric and magnetic fields. He also learns something of using the instrument, how the focussing and intensity controls interact, etc. Something of the techniques of measurement and their pitfalls are taught: the d'Arsonval instrument as ammeter, voltmeter, and ohmmeter, bridges, pen recorder, oscilloscope, the d'Arsonval instrument with a rectifier, etc. When the student gets to circuit theory he uses sine-wave and square-wave generators to observe the transient and steady-state behavior of networks. Anomalies, predicted by field theory but not by the lumped model, are observed. In one of the most popular experiments the student is required to deduce the configuration of elements in one of a set of 2-terminal black boxes, using any measuring equipment that he thinks will help him toward the answer. There are two experiments on d-c machinery, and they are surprisingly popular, perhaps because the students enjoy their mastery over a powerful machine but do not have enough time with it to become bored. Transformers, with emphasis on frequency response, and some basic experiments on polyphase networks, end the year's work.

Alongside the introductory course in electrical engineering the sophomore student finishes his sequence in the elementary calculus and takes a semester of differential equations. The second semester contains a physics course in electromagnetic theory that uses vector calculus and carries the student further into the mathematical aspects of the subject.

At this point the student is far enough advanced with ideas and techniques that his succeeding courses can proceed rapidly. As a junior, he has a semester of engineering analysis, a semester of energy conversion and control, and a year of communication systems and electronic circuits. The junior course in communication presents the elementary system aspects of the subject before the circuitry since it is believed that the question of "why" should be answered before "how." First, simple communication theory and signal theory are presented. After the student has an idea of what he is trying to do, vacuum-tube and transistor circuits are introduced as convenient means of accomplishing the desired results. The classroom treatment continues to discuss theory and the models of physical apparatus, while the laboratory shows the practical features, the limitations of the models, and the techniques of measurement, and emphasizes the use of judgment in design. Here the student learns by hard experience the effects of the wrong decimal point, the ground loop, and leads that are too long. To save time, pegboards are generally used to hold individually mounted components and leads with friction connectors are used for wiring. But sometimes the student must deal with manufactured gear and must read its diagrams and locate its parts, and sometimes he must make a real breadboard and solder the joints well.

Together with the junior courses in electrical engineering, the student takes a mathematics course in functions of a complex variable and a semester of modern physics. He often elects additional physics courses in atomic and electromagnetic theory. Many students take analytical mechanics in the physics department rather than in mechanical engineering. Thermodynamics, also, can be taken in either department.

This is all basic preparation for the senior year, where the student takes a more advanced course in electronic circuits and chooses among physical electronics, feedback control, microwave devices and applications, signal and communication theory, switching theory and logical design, digital computation, further courses in mathematics and physics, and independent work.

The program is a challenging one, no less in the sophomore year than later. There were some who originally feared that with an introductory course of such an ambitious nature the department would lose students, but they underestimated the motivating effect of a challenge. Students, instead, transferred into the department. A common statement was to the effect, "I hear that your department is difficult but that your students learn a lot. I want to join up with you." We are pleased at the increasing number of students from physics, mathematics, chemistry, and even economics who elect our courses. On the basis of all evidence, the experiment has been successful.

A New Method of Space Exploration by Radar, Using Incoherent Scattering

T. LAASPERE
NONMEMBER AIEE

Summary: Incoherent scattering is discussed in the context of the phenomenon of scattering of radio waves by spatial fluctuations in electron density. A review is given of both theoretical and experimental results. Emphasis is placed on the ideas involved in measuring properties of the upper ionosphere, such as the electron density, temperature, electron drifts, and the strength and direction of the magnetic field.

RETURNED RADIO WAVES from the lower ionosphere were used to study this region a long time before it was penetrated by the first research rocket. Since 1951, radio scientists have obtained information about the highest regions of the ionosphere as well. This is done by analyzing low-frequency electromagnetic waves that are generated by lightning strokes and travel from one hemisphere to the other along the lines of force of the earth's magnetic field in the "whistler mode." Very recently a method suitable for studying almost the whole ionosphere, out to a distance of 1 earth-radius or more, has been developed. This new method of space research is based on the phenomenon called incoherent scattering of radio waves. This paper discusses the fundamental ideas and techniques of using this form of scattering to measure some of the basic properties of the ionosphere. Before this is done, however, a short discussion is given on the phenomenon of radio wave scattering in general.

Radio Wave Scattering in the Ionosphere

About 1948, it was realized in connection with the propagation of high-frequency waves through the lower atmosphere (the troposphere) that the strength of the signal beyond the horizon was consistently too strong to be explained in terms of such concepts as diffraction and propagation in an atmospheric waveguide (duct). For instance, the interference of distant frequency-modulation-stations was found to be a much more serious problem than expected. The explanation was advanced that these relatively strong signals were the result of a process of scattering, in which some of the energy beamed in the forward direction was spread into other directions by irregularities in the spatial distribution of the dielectric constant. This explanation stimulated research in order to detect scattering in the ionosphere. Such a phenomenon was actually discovered in the lower ionosphere in about 1950.

In an ionized gas, radio waves are scattered by spatial fluctuations of electron density of scale comparable to the radio wavelength. (The word "scale" has a definite meaning in the theory of radio wave scattering. For the purpose of this paper, it is sufficient to interpret "scale" as being roughly equivalent to "size.") A radio wave scattered by irregularities in electron density can be detected if the intensity of the irregularities exceeds a certain minimum, determined by such considerations as the geometry of the setup, radio wavelength, transmitted power, effective antenna area, receiver sensitivity, and the intensity of any kind of interfering radiation.

The frequencies used are so high that no reflection of the wave results; thus, if the ionosphere did not contain any irregularities in electron density, all of the energy of the wave would pass unperturbed through the ionosphere into interplanetary space.

The fundamental idea behind all research using radio wave scattering is that—since the characteristics of the irregularities, and thus also of the scattered waves, are determined by some of the basic properties of the region in which scattering takes place—such properties may be deduced from an analysis of the scattered waves.

The phenomenon of radio wave scattering has received much attention with the recognition of its importance. This has lead not only to new techniques of long-distance communication, but also to some fruitful and novel ways of ionospheric research. One new method involves the illumination of auroral ionization by a powerful beam of radio energy, some of which is returned by the process of scattering. Another similarly investigates the ionized trails left by the multitude of small meteors that penetrate the earth's atmosphere each instant.

Until recently, the phenomena involving scattering of radio waves in the ionosphere were thought to be confined largely to the region below the maximum electron density, which occurs at about 300 km (kilometers) above the earth's surface. For instance, the irregularities associated with aurorae are usually observed at heights much below 300 km, although they have been found as high as that.[1]

Also, the rather intense irregularities, which have been detected in forward-scatter experiments at 30 to 100 mc (megacycles), exist at the relatively low level of about 85 km.[2] The latter irregularities may be produced by a process in which gradients of electron density are disturbed by turbulent motions in the neutral atmospheric gas. Such turbulent motions can still exist at sufficiently small scales at this height, since here the particle mean free paths are still quite small. At higher levels of the lower ionosphere (200 to 300 km), particle mean-free paths become of the order of hundreds of meters, and one might think that no irregularities of the order of meters can exist there. If an irregularity of the order of a meter were created there by an external agent in the neutral atmospheric gas, the irregularity would almost immediately diffuse. However, when the mean free paths become large, one must take into account that the earth's magnetic field begins to affect the behavior of charged particles.

In the presence of a magnetic field, the motion of charged particles in directions perpendicular to the field is restricted to distances comparable to their radii of gyration (although motion along the field is not affected). If this restriction on the particle motion is kept in mind, it is perhaps not surprising that sometimes, even in the range from 200 to 300 km, irregularities in electron density appear to exist with scales only of the order of meters in directions perpendicular to the field.[3] As might be expected, these irregularities appear to be elongated in the direction of the field. How they are produced is, at present, not well understood. The process may involve the action of irregular winds in the presence of the magnetic field, electrostatic effects, or other conditions.

Paper 61-1048, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted July 26, 1961; made available for printing September 15, 1961.

T. LAASPERE is with the Thayer School of Engineering, Dartmouth College, Hanover, N. H.

Incoherent Scattering

With the publication of a paper by W. E. Gordon of Cornell University in 1958, it became generally recognized that a certain kind of scattering must occur at all ionospheric heights at all times, regardless of whether the radio wavelength is large or small compared to the electron mean free path or to the radius of gyration.[4] The irregularities in electron density which support this kind of scattering are created by the electrons themselves in the course of their thermal motion. The intensity of such thermal irregularities, and thus also the intensity of the scattered signal to which they give rise, is several orders of magnitude weaker than the intensity of the irregularities which have been utilized previously in ionospheric research.

Although extremely weak, scattering from thermal irregularities shows great potentiality as a tool for exploring the ionosphere above 100 km, and has already yielded some rather interesting results. This kind of scattering, in principle, could also be used to transmit information over large distances. Its existence, incidentally, can be interpreted to mean that, for radar detection problems, an ionized medium has an inherent noise associated with it.

ELECTRON DENSITY

This new form of radio wave scattering is called incoherent scattering because, for an ionized gas in thermal equilibrium, a nearly correct prediction for the scattered power is obtained if the wavelets radiated by the individual electrons under the influence of the incident field are added directly in terms of power, i.e., incoherently.

The requirement for such addition of powers is that the spatial distribution of electrons be completely random over distances at least as large as a wavelength, which should nearly always be approximately the case in the upper ionosphere at wavelengths used in radio wave scattering. The scattered power is then directly proportional to electron density, which can be deduced with relative ease from the strength of the received echo. The methods of incoherent scattering should prove invaluable in investigating a whole range of such phenomena of the upper ionosphere as are accompanied by adequate changes in electron density.

TEMPERATURE

Time variations of the spatial fluctuations in electron density that give rise to scattering cause a monochromatic transmitted wave to be received as a spectrum of frequencies. (This is somewhat similar to what would be observed if a transmitted wave were reflected by a large number of moving objects.) In incoherent scattering, these time variations are caused by thermal motions of the particles of the medium. Hence, in the absence of electron collisions, it is natural to expect the frequency spread to be directly proportional in amount to the mean thermal velocity of the ionized particles, which, by the kinetic theory of gases, is in turn proportional to the square root of the temperature. The temperature of the medium can, in fact, be deduced from the width of the received spectrum, if the complications that arise from the existence of Coulomb forces of interaction between the charged particles are taken into account. The collective effect of these forces is to couple the dynamics of the fluctuations in electron density to those in the density of the positive ions, if the scale of the fluctuations is larger than a critical length—the so-called Debye length (of the order of a centimeter at a height of 1,000 km).

The coupling between the irregularities in electron and ion densities occurs because an ionized medium tends to remain electrically neutral over distances larger than the Debye length. Under conditions of thermal equilibrium, therefore, large-scale irregularities in electron density cannot exist unless they are accompanied by similar, neutralizing irregularities in the density of the positive ions. In the presence of this coupling, most changes in the distribution of electron density occur at a rate determined by thermal velocities of the relatively massive, slow-moving, positive ions, whereas in the absence of coupling (i.e., at scales smaller than the Debye length) these changes occur at the rate which is determined by the thermal velocities of the light, fast-moving electrons.

Since scattering is caused by spatial fluctuations in electron density whose scale is comparable to the wavelength, a relatively narrow spectrum of backscatter should be expected, if the radio wavelength is larger than the Debye length, but a broad spectrum if the reverse is the case. A theory of incoherent scattering in the presence of this coupling has been developed as a result of a number of studies.[5–10] An interesting prediction of the theory is that, at wavelengths larger than the Debye length, the spectrum of incoherent backscatter from the ionosphere should, in general, exhibit a rather flat top with a slight dip at the center frequency.

DRIFTS OF IONIZATION

Drifts of ionized clouds can be deduced from the displacement of the whole spectrum of scattering that such motions introduce. Such a displacement of the spectrum can be thought to be caused by the mechanism of Doppler shifts, in much the same way a radar echo from a moving target undergoes a change of frequency.

MAGNETIC FIELD

The presence of a constant magnetic field affects the motion of a charged particle by causing it to gyrate about a magnetic line of force at a frequency directly proportional to the strength of the field and inversely proportional to the mass of the particle. If all effects of electrostatic forces on the motion of an electron can be neglected, the effect of the earth's magnetic field on the spectrum of incoherent backscatter can be determined by an approach that is relatively simple.[11]

The radiation field of a typical electron is analyzed and found to involve both phase modulation and Doppler shift. Both are caused by the component of the electron's motion in the radar line of sight—phase modulation by the electron's gyration, and Doppler shift by its drift. The electrons can drift only parallel to the magnetic lines of force, so that the line-of-sight drift velocity, and thus also the Doppler shift, is zero for each electron when the radar beam is directed perpendicular to the magnetic lines of force. In this case, a line spectrum results, with the separation of the lines equal to the frequency of gyration (gyromagnetic frequency of the electrons).

If the radar beam is not perpendicular to the lines of force, the line-of-sight drift velocities have a spread about zero, leading to a similar spread in the Doppler shifts. Both spreads are directly proportional to the cosine of the angle between the radar beam and the lines of force. Therefore, as the radar beam is swung away from perpendicularity, the spectral lines broaden. As this broadening approaches the electron gyromagnetic frequency (separation of the lines in the line spectrum), the whole spectrum becomes smeared. At larger deviations from perpendicularity, the effects of the magnetic field cannot be observed.

The approach just outlined can be used only for radar wavelengths smaller than the Debye length, since only over spatial distances smaller than this length can electron motion be thought to proceed unaffected by the presence of all other charged particles. In incoherent scattering, large electron deflections (collisions) can always be neglected in the upper ionosphere. However, small perturbations in electron motion, caused by the interaction of an electron with the multitude of other charged particles through the long-range Coulomb force, can be neglected only for distances smaller than the Debye length.

A collective effect of these perturbations has already been stated in the section on temperature. The perturbations lead to the positive ions controlling the dynamics of those fluctuations in electron density which have a scale larger than the Debye length. Because of this control, it is reasonable to expect that, at large radar wavelengths, the ion gyromagnetic frequency will play a role somewhat similar to the electron gyromagnetic frequency at small wavelengths.[6] This expectation is confirmed by the results of a number of theoretical analyses.[12-18] These results show that the phenomenon of incoherent scattering can, in principle, be used to determine not only the direction and strength of the magnetic field, but also the masses of the positive ions present. The magnetic field will also enter the problem by affecting the radio wave polarization (Faraday rotation).

Experimental Techniques

An incoherent-scatter experiment can be performed either with a radar, which operates as a receiver when not transmitting, or with a setup in which transmitting and receiving facilities are separate and far removed from each other.

With radar, a long train of relatively short pulses is transmitted, and the scattered signal is observed in the intervening gaps or spaces. Since irregularities in electron density exist in all of the ionosphere, each pulse of transmitted energy gives rise to a scattered signal during its whole journey through the ionosphere. Eventually the signal of backscatter from a particular pulse will, however, become too weak to be of any concern. In incoherent scattering, there is a factor of one over the distance squared, tending to reduce the signal received from a pulse, and, above the height of the maximum electron density of the ionosphere, the pulse will also travel through a region of rapidly decreasing electron density.

Thus, if the interval between successive pulses is kept sufficiently long, it is reasonable to consider that the received signal results from scattering by a single pulse. (If pulses are transmitted at the rate of 100 per second, contributions to the signal are received from volumes in space which are separated by 1,500 km from center to center.)

The size of the volume contributing to the received signal at each instant is important, since this size determines the space resolution that can be achieved. Simple considerations show that, if the leading edge of a pulse of duration τ is transmitted at the time $t=0$, the backscatter signal at the radar at the time t arrives from the range interval that is located between $c(t-\tau)/2$ and $ct/2$, where c is the velocity of propagation of the pulse (approximately the velocity of light in free space at frequencies employed in incoherent scattering). That is, at any instant, an incoherent-scatter radar receives signals from a volume whose dimension in the direction away from the radar is equal to half the pulse length in space (150 meters for a microsecond pulse), and whose extent sidewise is determined by the antenna beam.

Also noteworthy is the fact that whereas at time t the pulse itself is at a distance ct, the contributing volume is at $ct/2$, so that the distance to the volume in space from which the scattered signal is received increased at the rate $c/2$. It follows that the signal of backscatter of a time interval Δt originates in a total range interval $c\ \tau/2+c\ \Delta t/2$. However, the central portion of this range interval is more heavily weighted in the sense that it contributes to the received signal during a larger fraction of the time interval Δt than the regions nearest and farthest from the radar.

If the transmitting and the receiving facilities are far apart, the scattered signal is received only from a volume common to the beams of both transmitting and receiving antennas. In such cases, since pulse techniques are not necessary to define a scattering volume, the transmitter could be made to transmit continuously. This may be a highly desirable feature in experiments involving measurements on the spectrum of scattering, since the technique of using pulses introduces some complications into such measurements.

Experimental Results

With the radar techniques described in the previous section, the existence of incoherent scattering of radio waves was confirmed in the fall of 1958 by a group of engineers and scientists of the National Bureau of Standards, led by K. L. Bowles.[5] From the strength of the received echo, Bowles was able to trace out the profile of electron density versus height in the approximate range between 100 to 600 km. The frequency used in that experiment was 40.92 mc, and the antenna consisted of an array of 1,024 half-wave dipoles.

Observations of incoherent backscatter were subsequently extended at Lincoln Laboratory's Millstone Hill radar site, Lexington, Mass., to include information on frequency distribution of the scattered power.[19] The Millstone Hill results were obtained at 440 mc and indicate an ionospheric temperature, deduced from the width of the flat-topped spectrum predicted by the theory of about 1,000 degrees Kelvin or 1,340 degrees Fahrenheit. More recently, some effects of the earth's magnetic field in incoherent backscatter have been observed with a radar located on the British West Indies island of Trinidad.[20]

Radars Designed to Make Use of Incoherent Scattering

While some use has already been made of this new method of space research, its full utilization awaits the completion of two extremely powerful and sensitive radar systems now under construction. The radars were designed primarily to use the phenomenon of incoherent scattering for upper atmospheric research.

One radar site is near Arecibo, Puerto Rico.[21] The radar reflector will be shaped to resemble the section of a sphere; will have an 870-foot radius and an aperture of 1,000-foot diameter; and will be fitted into a bowl-shaped limestone sinkhole. By moving the antenna feed, this radar beam (1/6 degree at the operating frequency of 430 mc) can be directed anywhere within a 40-degree cone centered at the vertical axis. The design and construction of this facility is under the direction of W. E. Gordon of Cornell University's Center for Radiophysics and Space Research. The whole project is sponsored by the Advanced Research Projects Agency of the Department of Defense through the Air Force Cambridge Research Center.

The other incoherent-scatter radar is being built near Lima, Peru, by Bowles' group from the National Bureau of Standards. This radar will have an antenna consisting of some 18,000 dipoles with a very limited beamswing capability. The facility is located near the magnetic equator, however, so that the radar can operate with its beam perpendicular to the magnetic lines of force. Since the effects introduced by the earth's magnetic field are predicted to show up best in only a small range of angles near perpendicularity, it is possible that some effects will be detectable with this radar.

In addition to the two radars just described, an antenna with a 150-foot aperture is nearing completion in the United States at Stanford University, Calif., it, too, probably will be used for research in incoherent scattering. Further work in this field may also be continued at Millstone Hill and Trinidad.

Conclusions

The phenomenon of incoherent scattering of radio waves has been discussed from the viewpoint of its use in ionospheric research. The number of electrons in the scattering volume can be deduced with relative ease from the intensity of scattered radiation, the temperature from the width of the spectrum, and electron drifts from a shift of the whole spectrum. Under certain conditions, the spectrum of scattering will also be affected by the earth's magnetic field, with the result that, in principle, both its magnitude and direction could be determined. It should also be possible to use incoherent scattering to study the composition of the ionized gas.

Incoherent scattering will surely be an important addition to the ways of exploiting various wave-propagation phenomena for the purpose of space research. It is certain that, even in this age of rockets and satellites, the ground-based radio engineer will make important contributions to understanding the vast ionized medium that surrounds the earth.

References

1. High-Altitude 106.1-MC Radio Echoes from Auroral Ionization Detected at a Geomagnetic Latitude of 43°, **J. C. Schlobohm, R. L. Leadabrand, R. B. Dyce, L. T. Dolphin, M. R. Berg.** *Journal of Geophysical Research*, Richmond, Va., vol. 64, 1959, pp. 1191–96.

2. Radio Scattering in the Lower Ionosphere, **H. G. Booker.** *Ibid.*, pp. 2164–77.

3. Evidence of Elongated Irregularities in the Ionosphere, **B. Nichols.** *Ibid.*, pp. 2200–01.

4. Incoherent Scattering of Radio Waves by Free Electrons with Applications to Space Exploration by Radar, **W. E. Gordon.** *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 46, 1958, pp. 1824–29.

5. Incoherent Scattering by Free Electrons as a Technique for Studying the Ionosphere and Exosphere: Observations and Theoretical Considerations, **K. L. Bowles.** *Journal of Research*, National Bureau of Standards, Washington, D. C., vol. 65D, 1961, pp. 1–14; also, *Report no. 6070*, National Bureau of Standards, Sept. 18, 1959.

6. A Theory of Incoherent Scattering of Radio Waves by a Plasma, **J. P. Dougherty, D. T. Farley.** *Proceedings*, Royal Society, London, England, vol. 259, series A, 1960, pp. 79–99.

7. Scattering of Radio Waves by an Ionized Gas in Thermal Equilibrium, **J. A. Fejer.** *Canadian Journal of Physics*, Ottawa, Ont., Canada, vol. 38, 1960, pp. 1114–33.

8. Long-Range Interactions in Ionized Gases in Thermal Equilibrium, **F. D. Kahn.** *Astrophysical Journal*, Chicago, Ill., vol. 129, 1959, pp. 205–16.

9. Scattering of Radio Waves from a Nondegenerate Ionized Gas, **J. Renau.** *Journal of Geophysical Research*, vol. 65, 1960, pp. 3631–40.

10. Electron Density Fluctuations in a Plasma, **E. E. Salpeter.** *The Physical Review*, New York, N. Y., vol. 120, 1960, pp. 1528–35; also, *Research Report no. RS 18*, Center for Radiophysics and Space Research, Cornell University, Ithaca, N. Y., Aug. 1, 1960.

11. On the Effect of a Magnetic Field on the Spectrum of Incoherent Scattering, **T. Laaspere.** *Journal of Geophysical Research*, vol. 65, 1960, pp. 3955–59; also, *Research Report no. RS 15*, Center for Radiophysics and Space Research, July 15, 1960.

12. Fluctuations in a Multicomponent Plasma, **O. Buneman.** *Journal of Geophysical Research*, vol. 66, 1961, pp. 1978–79.

13. A Theory of Incoherent Scattering of Radio Waves by a Plasma in a Uniform Magnetic Field, **D. T. Farley, J. P. Dougherty, D. W. Barron.** *Report no. US 21*, Research Laboratory of Electronics, Chalmers University of Technology, Gothenburg, Sweden, 1961.

14. Scattering of Radio Waves by an Ionized Gas in Thermal Equilibrium in the Presence of a Uniform Magnetic Field, **J. A. Fejer.** *Canadian Journal of Physics*, vol. 39, 1961, pp. 716–40.

15. Density Fluctuations in a Plasma in a Magnetic Field with Applications to the Ionosphere, **T. Hagfors.** *Journal of Geophysical Research*, vol. 66, 1961, pp. 1699–1712; also, *Scientific Report no. 1*, Radioscience Laboratory, Stanford University, Stanford, Calif., Dec. 5, 1960.

16. The Spectrum and Total Intensity of Electromagnetic Waves Scattered from an Ionized Gas in Thermal Equilibrium in the Presence of a Static Quasi-Uniform Magnetic Field, **J. Renau, H. Camnitz, W. Flood.** *Report no. UB-1483-P-1*, Cornell Aeronautical Laboratory, Inc., Buffalo, N. Y., Mar. 18, 1961.

17. Effect of the Magnetic Field in Ionospheric Backscatter, **E. E. Salpeter.** *Journal of Geophysical Research*, vol. 66, 1961, pp. 982–84.

18. Plasma Density Fluctuations in a Magnetic Field, **E. E. Salpeter.** *The Physical Review*, vol. 122, 1961, pp. 1663–74.

19. Some Characteristics of Ionospheric Backscatter Observed at 440 MC/S, **V. C. Pineo, L. G. Kraft, H. W. Briscoe.** *Journal of Geophysical Research*, vol. 65, 1960, pp. 2629–33.

20. The Effect of Faraday Rotation on Incoherent Backscatter Observations, **G. H. Millman, A. J. Moceyunas, A. E. Sanders, R. F. Wyrick.** *Journal of Geophysical Research*, vol. 66, 1961, pp. 1564–67.

21. The Design and Capabilities of an Ionospheric Radar Probe, **W. E. Gordon, L. M. LaLonde.** *Transactions*, Professional Group on Antennas and Propagation, Institute of Radio Engineers, New York, N. Y., vol. AP-9, no. 1, Jan. 1961, pp. 17–22.

A New Billing and Collection System Using Teletypewriter Techniques

E. W. HENCHES
MEMBER AIEE

Summary: Preparation each month of customers' telephone bills, and processing the bills after payment, are tremendous jobs. A new billing and collection procedure has been developed which combines a standard punched-card accounting system with recently developed teletypewriter and optical scanning equipment. This extended mechanization, now in operation, has made possible more efficient methods and has effected substantial savings. It can be adapted to other machine billing systems.

A SUBTITLE for this paper might well be: "The Automation of the Automated." For herein is a record of how accounting and engineering department co-operation and machine suppliers' assistance resulted in further economies in a generally accepted and efficiently operating data-processing system.

Large and legion are the accounting, billing, and collecting problems in the telephone industry, which were studied for the purpose of giving better and more accurate customer billing service and of enabling the revenue accounting forces to cope with the continuing growth in customer demands.

The New York Telephone Company began using IBM (International Business Machines Corporation) punched-card accounting machines a number of years ago for payroll work, pay checks, and several other plant accounting operations. Customer billing, however, continued to be the product of a variety of machines, some of which were semiautomatic although most were manually operated.

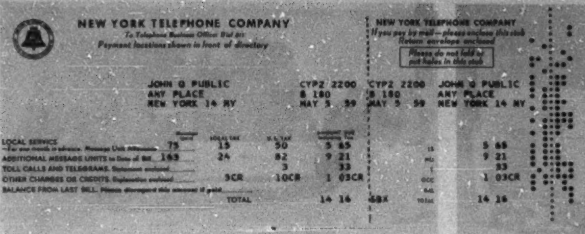
During 1955, the company decided to use automatic data-processing machines for preparing customers' bills, and the first revenue accounting office thus employing IBM punched-card equipment began operations in October 1956.

This system efficiently prepared and printed the customers' bills, but the ac-

---

Paper 61-1029, recommended by the AIEE Telegraph Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted June 12, 1961; made available for printing August 8, 1961.

E. W. Henches is with the New York Telephone Company, New York, N. Y.



NEW YORK TELEPHONE COMPANY
To Telephone Business Office: [illegible]
Payment locations shown in front of directory

NEW YORK TELEPHONE COMPANY
If you pay by mail—please enclose this stub
Return envelope enclosed
Please do not fold or put holes in this stub

JOHN Q PUBLIC
ANY PLACE
NEW YORK 14 NY

CYPZ 2200
B 180
MAY 5 59

		LOCAL TAX	U.S. TAX			
LOCAL SERVICE—For one month in advance. Message Unit Allowance	75	15	50	5 65	LS	5 65
ADDITIONAL MESSAGE UNITS to Date of Bill	163	24	82	9 21	MU	9 21
TOLL CALLS AND TELEGRAMS. Statement enclosed			3	33	T	33
OTHER CHARGES OR CREDITS. Explanation enclosed		3CR	10CR	1 03CR	OCC	1 03CR
BALANCE FROM LAST BILL. Please disregard this amount if paid					BAL	
TOTAL				14 16	6BX TOTAL	14 16

Fig. 1. Perforated bill

counting people were not quite satisfied with payment recording procedures. Each bill stub returned with a customer's payment had to be key-punched for the record, and other related work required much clerical time.

Believing that further mechanization was possible, the accounting people who had introduced machine equipment into the billing job consulted with their engineering co-ordinates.

The Idea

The accounting project had studied many payment-handling methods—some on trial, some proposed—most of which centered around some form of punched card to be sent to the customer as a bill and returned, in whole or in part, with the payment. A system that utilized the paper stub of the customer's bill, with pertinent payment data edge-punched into it in Teletype code, was considered most advantageous because of the following:

1. The edge-punching can be done on a single pass through the accounting machine at the time the bill is printed.
2. Payment information can be transmitted over teletypewriter circuits to distant points where such data are needed as quickly as possible.
3. Perforated stubs can be used as source material for the automatic preparation of other records necessary to the payment operations.
4. The edge-punched stub system is compatible with large computer operation.
5. The system is marketable to our customers who have similar billing problems.

In 1956, the New York Telephone Company accounting and engineering people, along with engineers from IBM and the Teletype Corporation, began developing the necessary equipment to produce the edge-punched paper-bill stub.

By January 1957, telephone company engineers had developed a prototype card reader-translator for converting accounting-machine signals into Teletype signals, and the IBM engineers had made the necessary modifications in their equipment to insure joint operation. The Teletype engineers developed a stub perforator that operated under control of the card reader-translator to edge-punch the stub portion of the standard customer bill. They had also developed a stub reader required to read the perforated stubs. This equipment will be described later.

The Customer Bill

A customer's bill, as printed on an accounting machine and perforated in Teletype code by the equipment just mentioned, is shown in Fig. 1.

Each bill is the end product of several revenue-accounting operations, involving the collection, computation, and tabulation of much data. For this report, details concerning each operation are unnecessary. Suffice it to say that a group of punched cards, called a DECK, must be prepared and assembled for each of the accounts.

Identifying information—customer's name, address, and telephone number—is punched in one card. The basic charges and call allowances for his telephone equipment are punched in another. Separate cards show extra charges where more than the allowable message units for local or suburban calls have been made, as well as all long-distance charges for the billing period. If there are installation charges or credits, or an unpaid balance from prior bills, separate cards for these items are also prepared. Between 25 and 27 Teletype characters, depending upon the local revenue accounting office's requirement, are edge-punched into the stub. These include the telephone number, the total amount of the bill, and several identifying and control signal characters needed for later processing operations. No shifting from upper to lower case characters is required since only upper case figures and symbols on the Teletype keyboard are used.

Bill Preparation

The physical arrangement of printing-and-perforating equipment, together with the card movement, are shown in Fig. 2.

When the billing day arrives, the previously prepared DECK of cards for each customer is run through an accounting machine, which prints each item of the bill and computes and prints the total amount due. As a simultaneous side operation, the accounting machine controls an associated summary punch, which punches a new card called ACCOUNTS RECEIVABLE. The ACCOUNTS RECEIVABLE card, when punched, is mechanically moved into the adjacent card reader-translator, where the punched-card language is read and translated into Teletype language.

The stub perforator, mounted atop the accounting machine, perforates the bill stub. The ACCOUNTS RECEIVABLE card, having been read, is deposited in a card stacker located in the side of the card reader-translator. The cards are then filed locally in the revenue accounting office until they are needed in processing the paid bills.

Bills are printed and perforated and ACCOUNTS RECEIVABLE cards are produced at a rate of 1,400 per hour by this machine combination.

The Machines

CARD READER-TRANSLATOR

A brief digression for a bit of history seems in order. The original card reader-translator was designed with standard telephone relays and was electrically coupled to the summary punch. In addition to the signal-translation feature, it controlled and synchronized the accounting machine and the stub perforator operation. The original reader-translator was reliable, accurate, and had an insignificant amount of maintenance "down time" in more than 3 years' service, and was not retired until the present standard model became available in 1960.

Once the breakthrough on the edge-

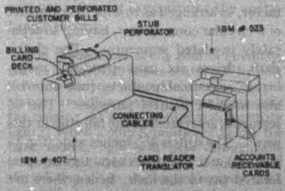


Fig. 2. Arrangement of printing and perforating equipment

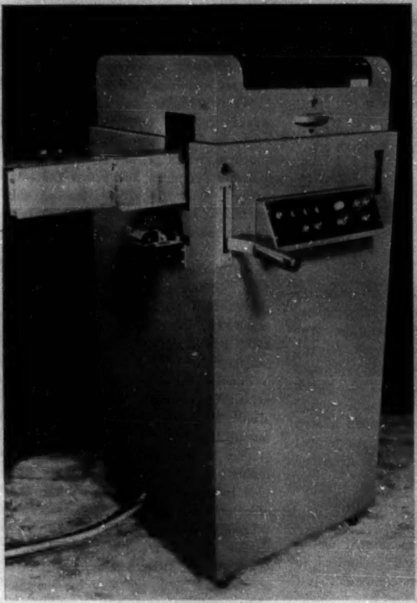


Fig. 3. Card reader-translator

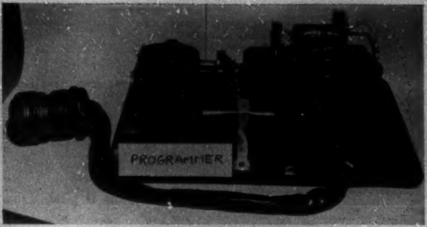


Fig. 4. Programmer unit

Fig. 5. Stub perforator

punched bill-stub method was made, the Teletype Corporation undertook the work of refining original models, and produced the aforementioned card reader-translator, which is now the New York Telephone Company standard.

This machine is shown in Fig. 3. It differs in concept from the original in that (1) it is not electrically connected to the summary punch or any other accounting machine equipment; and (2) it is fitted with a conveyor arm for mechanically transporting punched ACCOUNTS RECEIVABLE cards from the summary punch to a card-reading station within itself.

The principal control component, the programmer shown in Fig. 4, is a motor-driven, clutch-controlled, revolving contactor, so arranged as to wipe over the face of a circular commutator, having 35 separately insulated segments. The clutch shaft drives six cam-operated "make-break" electrical contacts (not shown), also used for control purposes. The 35 segments are related to the bill stub in this way: The stub is $3^1/_2$ inches wide (top to bottom) and Teletype characters are perforated ten to the inch; hence, there are 35 characters, each of which is associated with a programmer segment.

While there is room for edge-punching 35 characters across the $3^1/_2$-inch side of the bill (refer to Fig. 1), in practice the leading three and trailing three character spaces are coded as blanks. The remaining 29 spaces may be coded to punch fixed or variable characters, but only 27 are presently used, two spaces being reserved for possible future needs. Fixed characters are for permanent identification and machine-control characters. Variable characters, which are for telephone numbers, dollar amounts, and certain identification numbers or symbols, are controlled by codes punched in the ACCOUNTS RECEIVABLE card.

One rotation of the programmer contactor constitutes a machine cycle. During each cycle, a new card is read, and circuits—either through holes in the punched cards or directly from the segments of the programmer—are fed to input terminals of a diode matrix. The matrix has 32 input and 6 output terminals.

Each input terminal, when energized, causes, through solid-state circuitry within the matrix, a particular combination of signals in Teletype code to appear on 1 or more of the 5 output leads. The sixth output lead is used for parity checking. Each output terminal is cabled to a select magnet located in the stub perforator. These, in turn, control the punching operations.

The programmer, in addition to controlling the card reading, also controls the conveyor arm, which moves the card just read into the stacker and advances a new card into reading position for the next cycle. Circuitry is provided in the card reader-translator to insure that one hole, and only one hole, is punched in each column of the card that is read. Visual and audible alarms are provided to indicate machine malfunctions.

Since, as previously mentioned, the card reader-translator is not electrically connected to the summary punch, it is easily movable and can be placed adjacent to the summary punch when required. The conveyor arm readily fits into the output card stacker of the summary punch.

STUB PERFORATOR

The stub perforator is mounted on top of the accounting machine, as shown in Fig. 5, and is positioned to align with the paper-feed mechanism of that machine. It is controlled by the card translator.

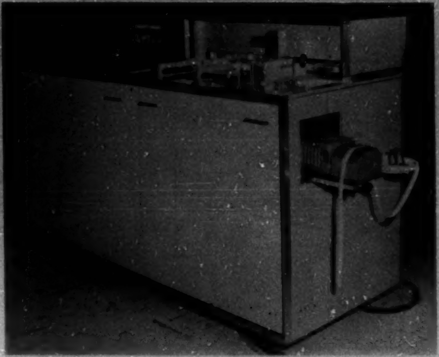


Fig. 6 (above). Stub reader

Fig. 7 (right). Optical scanner

The perforator is a sprocket-feed paper-moving device, combined with a 6-level perforating device. The paper feed is driven by a separate motor and is clutch-controlled. As mentioned before, the six selector magnets controlling the punching pins are connected by cable to the diode matrix in the card reader-translator.

This is an opportune time to discuss the lag in perforation of the bill stub. For space reasons, it is impossible to perforate and print the bill simultaneously. Actually, the punching is done 4 bill-printing cycles later. This delay is accomplished as the ACCOUNTS RECEIVABLE card advances, cycle by cycle, through the summary punch and along the conveyer arm of the card reader-translator until it reaches the reading station. By this time, the associated printed bill has been advanced by the paper feeds to the punching position.

The perforator is equipped with a tight-paper switch that is operated if, for some reason, bill printing is delayed and the normal amount of slack in the paper is taken up. The operation of the tight-paper switch will stop the programmer and thus halt the punching operation. This guards against tearing the paper if a delay occurs.

STUB READER

The stub reader, Fig. 6, became a necessity, once the perforated bill-stub system was developed. The reader is a specially modified tape transmitter, fitted with a paper-feed mechanism suitable for stub-handling. It is arranged to read 75 words per minute.

The first application of the stub reader is in producing the CASH card. This card is the revenue accounting-office's record that a bill is paid, in whole or in part. Customers are requested to return the cash-stub portion of the bill when making their payments. These stubs are fed into the stub reader, and its output operates a typing reperforator which produces punched tape. This tape is then put through a tape-to-punched-card converting machine, which produces a CASH card from the information originally punched in the bill stub.

The CASH cards thus produced are automatically sorted by telephone numbers. The ACCOUNTS RECEIVABLE cards, punched at the time the bill were printed, are removed from the files and processed through a collating machine with the CASH cards. The collator is programmed to match-or-merge the ACCOUNTS RECEIVABLE and CASH cards into batches of fully paid, partially paid, and unpaid accounts. The partially paid and unpaid ACCOUNTS RECEIVABLE cards become the balance-due file, and are processed at regular intervals to take care of payments which come in late.

The stub reader, typing reperforator, and tape-to-card conversion procedure, while capable of doing an automated job, invited further improvement. The accounting and engineering departments conducted trials of automatic feeders for the stub reader. Arrangements for direct control of the tape-to-card converter by the stub reader were also investigated. All effected some improvements, but none promised the speed offered by an electronic reading device that also was considered.

OPTICAL SCANNER

Negotiations were begun with a manufacturer of optical scanning equipment, and a scanning machine was developed for reading the perforated bill stubs and controlling a summary punch to produce CASH cards. The reading became 7 to 10 times faster than the stub reader method. The prototype of the scanner is shown in Fig. 7. The scanner is a self-contained unit. A control panel, containing all the keys and signal lamps required for its operation, runs along the top. The model shows the tape reader on a retractable mount in the side of the machine.

Present arrangements are to read bill stubs and control the punching of CASH cards at a rate of 100 per minute. Incorporated in the machine is a punched-tape reader, arranged to reach 600 words per minute. Speeds can be increased to 200 words per minute for the bill stubs and cards, and to 1,000 words per minute for the tape reader if this is later deemed desirable.

A simplified sketch of scanner operation is shown in Fig. 8. The perforated bill stubs, with holes uppermost, are placed in a feed stacker. A combined friction-belt–vacuum system picks the top bill off the stack and transports it past a reading station. An optics system, together with a revolving scanning disc, reads the perforations. Electronic circuitry will monitor and control the reading, translation, and card-punching. Various checking features, to insure a high degree of accuracy, are automatically incorporated.

A few excerpts from the manual on the scanner will assist in explaining the operational theory:

> "On the demand of the control circuitry, the stub transport passes stubs, one by one, before a narrow slit which serves as the scanner's reading station. Directly opposite the reading slit is mounted a high intensity light. When a stub is positioned before the light source, light will pass through the holes perforated in the stub. If you look at the reading slit as the optical scanner does, you will see a narrow rectangular area of bright light shining through each of the perforated holes. This image, projected

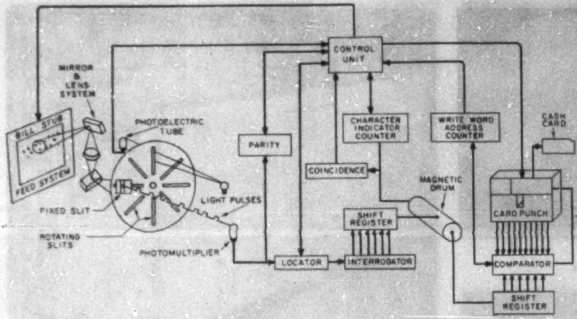


Fig. 8. Basic principles of scanner operation

by a prism and a series of lenses, is focused on a rotating scanner disc. Because of the bright light shining through the stub holes at the reading slit, the stub image focused on the scanner disc will have relatively light and dark areas. . . . The scanner disc is seven inches in diameter and rotates at a speed of 10,000 rpm."

The scanner disc actually has 30 radial slits. Behind the rotating disc is the fixed slit, through which pass the varying light intensities, depending upon whether a hole is being read or not.

"These varying intensities form a flow of pulses of light and these pulses are converted into electrical energy.

"A photomultiplier tube is employed for this task . . . From the photomultiplier unit the electronic pulse train is routed to an interrogating unit whose job it is to look at each incoming line of pulses to determine whether or not there is a hole punched in each of the six perforated lines in the bill stub. Having determined this, the interrogator sends this information to a shift register and then to a magnetic drum for storage. The shift register acts as a buffer. . ."

A buffer is needed to delay the recording on the magnetic drum until all signals for one character have been received. Then they are quickly recorded on the magnetic drum, revolving at 9,400 rpm so that the pulses are stored in a small area on the drum surface.

Stored information, which is in 6-level Teletype code, must be converted to the IBM code before it can be used to punch the CASH card. The comparator unit performs this job; it has two inputs: (1) a group of six leads from the magnetic drum through a shift register, and (2) a group of ten leads from the summary punch. As the name implies, the comparator compares the two inputs.

The unit looks at the magnetic drum recordings ten times, once for each of the horizontal rows of punched cards that might be used. Any row might require punches in any of the vertical columns. For example, a recording on the drum for row zero might show six holes; the recording of the next row below might show four holes and so on down the card until row one, the last, is read. The comparator determines in which vertical column of each row the card is to be punched. The character indicator-counter assists in this operation.

The condition of this counter shows the radial position of the magnetic drum at any time. For each character recorded on the drum, a pulse is sent to the counter. Through this, a direct relationship between the spacing of recordings on the drum and between the holes punched in the columns of the rows is established, and the punching can be monitored. Control circuitry now sends a signal to the associated summary punch to begin its function.

Mechanical and electrical operations of the reading function are synchronized by means of a photoelectric tube. The revolving slots of the scanner disc, interrupting a beam of light focused on the tube, cause the tube to send a stream of precisely timed electrical pulses to the control unit for monitoring and timing.

Bill stubs are read, CASH cards punched, and correctly read stubs are transported to the accept stack. Those not satisfying the scanner's self-checking features are routed to the REJECT stack, which receives less than 1% of all stubs read.

Among the machine's self-checking features is one for parity. (The parity-check punching was done when the bill stub was originally perforated. A hole was punched in the sixth level if the number of holes in the other levels totaled an even number. It was used to verify the correct punching of the bill by the stub perforator.) The scanner also utilizes this check to insure that it reads an odd number of holes through the six levels. If it does not, the stub is rejected. Another checking feature is: If a stub is correctly read, but not recorded on the magnetic drum, the control circuit is signaled and the stub is sent to the reject stacker.

To verify that all characters read are recorded on the magnetic drum, a write-word-address counter unit is provided, and is programmed for the required number of characters read from the stub. If the correct number is not counted, or if a character has been added, the control circuit, on signal from the counter, will send the stub to the REJECT stack.

As an end-of-message signal, the Teletype code for the letter *Z* was arbitrarily chosen and is always punched as the last character in every stub. In the mechanized collection procedures, it acts as a control signal for the various component machines. The scanner checking circuitry will reject a stub not containing a *Z* in the proper location.

Reading of payment information from perforated tape is accomplished by the scanner's tape reader. Control-panel key operation disconnects the stub-reading circuitry and connects that of the reader. Using the tape-read information, the machine translates and controls operation of the summary punch.

Outstanding features of this scanner are: the friction-belt pneumatic-stub feed system, automatic alignment of the stub to insure a proper reading position, and self-checking and monitoring circuitry that provides a high degree of accuracy.

ADDING MACHINE AND MATRIX

One unique advantage of this billing and collection system is its ability to send information over teletypewriter circuits economically and with a minimum of machine language conversions.

On-the-spot records were required of the money amounts represented in payment information being transmitted to distant points by the stub readers. Again, the accounting and engineering people devised a machine operation to accomplish this and, with assistance from engineers of an adding machine manufacturer, the equipment arrangement shown in Fig. 9 was developed.

The key to this combination is a solid-state device matrix, electrically located between the stub-reader and adding machine. The matrix uses power transistors as switching devices for translating Teletype code into decimal information, suitable for operating the adding machine.

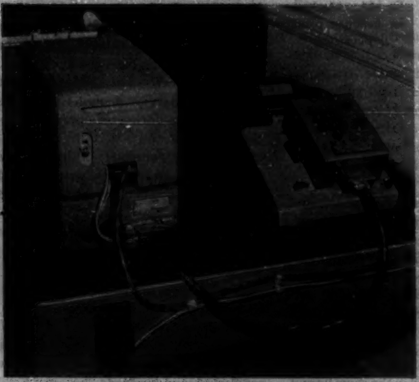


Fig. 9. (above). Stub reader, matrix, and adder combination

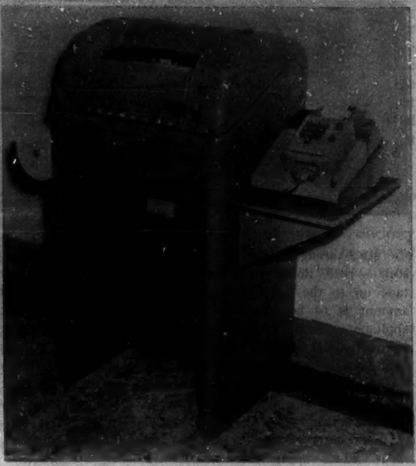


Fig. 10. (right). Typing reperforator and adder combination

This matrix is shown beneath the standard stub reader.

The adding machine is basically a 10-numeric-key unit, equipped with solenoid control, and arranged for add and nonadd features. Various fixed characters, punched into the bill stub, are arbitrarily used to control the adding machine.

As a stub is read, the matrix prevents registering any data until the line number is reached. By operating the proper solenoids, the line number is printed on paper tape as nonadd data. Subsequent information being read out of the stub is not printed until a control code is detected, just before the total amount of the bill is to be read. The adding machine now functions to register this amount. Reading the last character *Z* causes the adding machine to print the amount and enter it into its accumulator. When a batch of stubs is completed, the manual operation of a machine key will cause the printing of the total amount registered in the accumulator.

The transistor matrix receives its power from the stub reader. The adding machine is equipped with a separate power cord, which permits its use for regular office work when it is not busy with the stub reader in transmitting payment information.

Related Operations

TRANSMISSION OF PAYMENT INFORMATION

The equipment just described is used when customers pay their bills in person at telephone company public offices. The New York Telephone Company has not yet extended this method to agents now authorized to accept telephone bill payments, but this could be done if warranted by the volume of payments and other economic factors.

A stub reader–adding machine combination, Fig. 9, at a public office works with a typing reperforator–adding machine combination, shown in Fig. 10, located in the centralized remittance unit at a revenue accounting office. These machine groups can be connected by private lines, mechanized teletypewriter exchange systems, *81-D* Teletype systems, or by Dataphone. Again, the means selected would depend on economic considerations. So far, the New York Telephone Company applications have used private lines.

The information transmitted from the public office is received on a typing reperforator which produces punched tape. The associated adding machine will print and add the billing information in synchronism with the adding machine located in the remote public office. When transmission is completed, the clerks obtain the totals and verify the amounts, using the teletypewriter circuit for talking. The tape produced by the typing reperforator at the remittance unit is processed through a tape-to-card converter to produce CASH cards.

In addition to the public offices, the telephone company operates record offices where information pertaining to the nontechnical portion of the customer's service is kept. These records cover the equipment in use, credit data, toll bill details, account status, and similar items, which may be consulted when customers seek information from the record office. Every effort is made to keep the files accurate and up-to-date.

Stub readers in the centralized remittance units transmit payment information promptly to various record offices. Either bill stubs or punched tapes are read, and the payment data are reported on a teletypewriter page-printer at the record office, from whence they are promptly distributed to office representatives for use in updating customer-payment files.

SPECIAL PREPARATION OF PAYMENT INFORMATION

In several instances, the bill stub returned by the customer with his payment is not usable. Chief among these are cases where only partial payment is made. Since the stub is perforated to show the full amount due, standard processing must be changed to record the partial amount paid. The equipment to handle this operation, shown in Fig. 11, consists of a stub reader and a typing reperforator. A small locally developed applique unit, not shown, consisting of two or three relays, is wired between these machines. Basically, this is a character counter.

A stub involved in a partial payment has the paid amount written on it, after which a remittance clerk processes

Fig. 11. Machines for handling partial payments

the stub through a reader. When the stub is read, the reperforator produces tape up to the point where the total amount is to be punched. Here the applique circuitry functions to prevent further reading. The clerk then types in the partially paid amount as written on the stub. The applique circuitry continues the stub through the reader without further reading, and the tape produced by this operation is used to prepare CASH cards and to supply record offices with information.

In an insignificant number of cases, a bill stub is mutilated to such an extent that it cannot be correctly read, and sometimes stubs are never returned. A record of payments under such circumstances is punched into tape by the same machines which handle partial payments. Operation of a cutoff key makes the applique unit inoperative, and all characters are manually key-punched in the same format as on an automatically prepared stub.

The Schedule

This billing and collection system has been operating successfully in a number of revenue accounting centers in New York City. By June 1962, all New York Telephone Company centers, except in the Westchester area, will be placed on the perforated bill-stub basis just discussed.

The Westchester area placed a large computer in service in mid-1961. This unit handles a variety of operations, including billing and collections for Westchester. The computer project people have elected to use the punched-card method for present collections. It is believed that the perforated bill-stub method can be economically applied to this large computer's billing and collection procedures.

Conclusions

The described billing and collection system, through introduction of new equipment and methods, has increased efficiency. The methods are economical, giving a good rate of return. Fast and accurate information service is provided for the record office representative.

Through further mechanization, higher service and production objectives are possible.

A Buffer Store for Data Transmission

W. E. BAKER
ASSOCIATE MEMBER AIEE

F. E. FROEHLICH
ASSOCIATE MEMBER AIEE

Summary: A method is presented for connecting an information source emitting data at an arbitrarily fluctuating rate to a synchronous transmission system. Expressions are derived relating the fluctuating input data rate and the maximum transmission efficiency to the capacity of a buffer store necessary to establish this connection.

An acoustic delay line buffer store whose input and output are independent of each other is described. The buffer store has a capacity of 45 bits, operates with 2-μsec (microsecond) spacing between pulses, and employs a unique method of generating a 500-kc timing signal.

THE rapid growth of automatic data processing in industry has directed considerable attention toward the problem of transmitting digital information from one geographical location to another. Even though the sources and receivers of such digital information differ considerably many, if not most, eventually wish to communicate with each other using common carrier transmission media such as telephone or telegraph facilities.

For many years standard telegraph services have been available for record communication. While a variety of terminal equipment existed, the format and speed of transmission was confined to several standardized forms. Many of the newer demands for data transmission find standard telegraph speed too slow and the format too restrictive for efficient usage.

Digital information coming from different data sources may vary in code, format, and speed. The many varied applications of the transmitted information make it unlikely that a single code, format, or transmission speed satisfying most of the data sources can be achieved. Nevertheless, transmission of such data must be accomplished accurately and efficiently, and messages must be reproduced faithfully at the receiver. Thus, the question arises whether it is feasible to design a single data transmission system that can establish communication between any or at least many types of data sources and receivers. Such a terminal should be capable of accepting information covering a wide range of data rates and a wide range of digital codes and formats.

A major problem facing such a universal digital data system capable of allowing a large variety of data machines to communicate is synchronization. The recovery of a timing signal at the receiver is needed to regenerate the digital information; thus, synchronization is fundamental to binary digital transmission. The system described in this paper can be adapted to start-stop transmission but since the maximum efficiency ($66^{2}/_{3}\%$ in most cases) is quite low only continuously synchronous transmission is considered.

For the case of high-speed data transmission, that is, for cases approaching the Nyquist rate, synchronization is usually derived directly from mark-space transitions of the data signal or from frequencies specifically added for the purpose of synchronization.[1–3] Generally the recovered timing frequency is the result of averaging over many bit intervals. Such

---

Paper 61-1020, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted March 10, 1961; made available for printing August 29, 1961.

W. E. BAKER and F. E. FROEHLICH are both with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

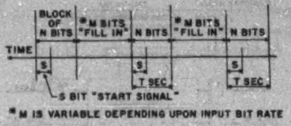


Fig. 1. Structure of the message

continuous synchronization recovery may be expected to give superior performance over a *start-stop* type for which synchronization depends on a single mark-space transition per character.

The precise timing required of continuous synchronizers for high-speed data transmission systems and the advantages gained from averaging the recovered timing wave over many bits usually results in synchronizers for which bit rates are not easily changed.[1,2] Thus, it has been found most advantageous for a data transmission system to be operated at a fixed bit rate between transmitting and receiving stations.

The desire to accept data from sources using a variety of codes, formats, and transmission speeds must be reconciled to the advantages of a fixed speed data transmission system. Even for cases where a particular fixed transmission bit rate chosen would be satisfactory, the accuracy in timing needed in feeding data to the transmitter is far beyond most normal clocking functions of the data sources. Thus, a problem of matching a fluctuating data source to a synchronous transmission system exists.

The solution to this problem lies in providing a buffer store for matching an asynchronous data source to a synchronous transmission system. Accordingly, this paper is most concerned with the insertion of such a buffer store between a data source having a fluctuating bit rate and a transmission system having a fixed synchronous bit rate. It is shown that a transmission system can be constructed that will transmit only at a single speed but that will be able to accept information over a continuum of bit rates totally independent of the code or format of the transmitted information.

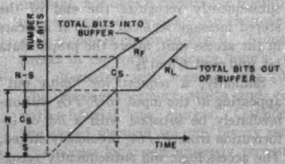


Fig. 2. Constraints on fastest input rate

Structure of the Message

In order to design a binary transmission system that is capable of accepting data from any source having an arbitrarily fluctuating bit rate, and to transmit the data at a synchronous rate, a buffer store should be inserted between the data source and the input to the transmission system. The buffer accepts fluctuating serialized binary data, stores the data temporarily, and, whenever a predetermined number of bits have been stored, releases the data in blocks of fixed length under control of the synchronous transmitting clock. No output appears from the buffer between transmitted blocks of information, but the synchronous transmitting clock continues operating so that the next block of information will start in the proper phase.

A similar buffer may be inserted between the synchronized receiver and the final destination, the data receiving machine. The receiving buffer is then able to restore the message to its original structure.

Assuming that the data source rate is always less than the transmission rate, although it may vary in any manner between the latter rate and a set lower rate, a message structure as shown in Fig. 1 may be formulated. Assume that the digits transmitted on the channel are divided into blocks of N bits and each block of N bits begins with a chosen starting signal consisting of S bits that is followed by a sequence of N-S information bits. Following the block of N bits are fill-in bits which carry no information. The number of these fill-in bits varies in accordance with the momentary input rate of the data source. These fill-in bits may consist of any prearranged sequence chosen so that bit synchronization at the receiving end of the transmission system may be maintained. In case mark-space transitions are not necessary for maintaining bit synch at the receiver, the fill-in bits may be chosen as a sequence of marks.

At the receiving end of the system, synchronization recovery is derived from the message block and fill-in bits. Further, knowing the structure of the starting signal and the number of information bits in each of the blocks of fixed length, the receiver is able to decode the message. In so doing no restrictions have been placed on the code or the format of the message to be transmitted from the data source.

A buffer store is necessary to create this message structure. The properties requisite to such a buffer are derived from the following considerations. If R_F bits per second and R_S bits per second are the fastest and slowest data source information rates, respectively, into the buffer store that a given system is expected to handle, and if R_L bits per second are the synchronous information rate on the transmission medium with the constraint that R_F is less than R_L, then the total fluctuation, δ, in the information input rate is defined to be

$$\delta=\frac{R_F-R_S}{R_F} \tag{1}$$

The maximum transmission efficiency $\epsilon_{(\max)}$ of the system that may be achieved occurs for a continuous input rate R_F, that is,

$$\epsilon_{(\max)}=\frac{R_F}{R_L} \tag{2}$$

The minimum transmission efficiency $\epsilon_{(\min)}$ correspondingly occurs for a continuous input rate R_S, that is,

$$\epsilon_{(\min)}=\frac{R_S}{R_L} \tag{3}$$

The constraints on such a system may be formulated in an indirect way. If the existence of such a system is assumed then C_S, S, R_L, and N are fixed, where C_S is the number of bits which must be accumulated in the buffer store before transmission of a new block of N bits can be started. What are the constraints on R_F and R_S, the fastest and slowest rates the system will accept? The fastest input rate is obtained when no fill-in bits are needed. This condition occurs when there are C_S bits remaining in the buffer at time, T, where

$$T=\frac{N}{R_L} \tag{4}$$

From Fig. 2 this may be seen as the following equation:

$$R_F=\frac{N-S}{N}R_L \tag{5}$$

The slowest input rate must be sufficient to insure that a complete block of N bits may be sent before the buffer is empty. This may be seen from Fig. 3 to be

$$R_S=R_L\frac{(N-S-C_S)}{N} \tag{6}$$

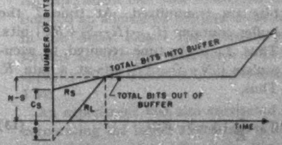


Fig. 3. Constraints on slowest input rate

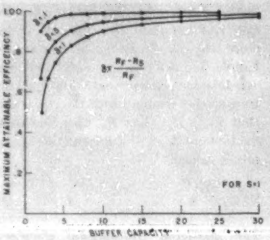


Fig. 4. Efficiency of transmission

Combining equations 5 and 6

$$C_S=S\frac{(R_F-R_S)}{(R_L-R_F)} \tag{7}$$

If, now, R_F and R_S are assumed to be given, along with R_L and S,

$$C_S=\left[S\frac{(R_F-R_S)}{(R_L-R_F)}\right] \tag{8}$$

where the brackets are used to denote the smallest integer which is greater than or equal to the quality within the brackets.

Solving equation 5 for N gives

$$N=\left[S\frac{R_L}{R_L-R_F}\right] \tag{9}$$

After C_S bits have been stored, the starting signal is sent, requiring S/R_L seconds for transmission. To insure that the buffer store does not overflow in this interval, additional storage, S', must be availabl eat the start of transmission of a block of N bits.

$$S'=\left[S\frac{R_F}{R_L}\right] \tag{10}$$

C, the capacity of the buffer, is then

$$C=\left[S\frac{(R_F-R_S)}{(R_L-R_F)}\right]+\left[S\frac{R_F}{R_L}\right] \tag{11}$$

Equation 11 gives the desired relationship between the size of the buffer store, the data source rate, and the line rate.

For a constant input rate, R, the number, M, of fill-in bits transmitted may also be determined:

$$M=R_LT_F \tag{12}$$

where T_F is the time during which fill-in bits are transmitted. At time T the buffer contains $[C_S-(N-S)+RT]$ bits. Thus, T_F is the time required to accumulate $[(N-S)-RT]$ bits at a rate R. Thus,

$$T_F=\frac{1}{R}\,[N-S-RT] \tag{13}$$

Combining equations 4, 9, and 13 yields

$$M=\left[S\frac{R_L}{R}\,\frac{(R_F-R)}{(R_L-R_F)}\right] \tag{14}$$

Equation 14 shows that the number of bits during the fill-in interval is adjusted according to the average data rate. As the average data input rate approaches R_F the number of fill-in bits decreases. For a steady input rate equal to the fastest permissible rate, R_F, no fill-in is employed.

From the previous equations more useful expressions, in terms of the constants of the system rather than particular bit rates, can be derived. Substituting equations 1, 2, and 3 the following expression for C_S is obtained:

$$C_S=\frac{(\epsilon_{(\max)}-\epsilon_{(\min)})S}{1-\epsilon_{(\max)}} \tag{15}$$

It should be noted from equations 2 and 5 that the required N is completely defined by the desired $\epsilon_{(\max)}$ and S, that is,

$$N=\frac{S}{1-\epsilon_{(\max)}} \tag{16}$$

The possible values of $\epsilon_{(\max)}$ and R_F are limited to those which are consistent with integer values for C_S, S, and N.

In Fig. 4 a family of curves showing the relations between $\epsilon_{(\max)}$, δ, and C is shown, assuming that the number of bits making up the starting signal equals unity, that is, $S=1$. It can be seen that the curves rise rapidly from a buffer capacity of 2 to approach a maximum efficiency of 75% or more with only 5 or 6 bits of storage capacity, regardless of the degree of fluctuation of the input rates.

A special case of the message structure discussed is formed by a transmission system using start-stop bit synchronization, such as standard telegraph. Here the maximum transmission efficiency, the transmission block length, and the starting signal are predetermined. A data source of fluctuating bit rate and arbitrary code or format can be connected to this system by using a buffer store of the capacity determined by equation 15.

The Buffer Store

For many cases of practical interest a buffer store of 5 or 6 bits storage capacity is able to satisfy the requirements specified in the previous sections. Buffer stores of such size having the proper characteristics can be designed.

There are, however, many other functions that a buffer store in a data terminal could perform when transmission is not taking place. Such uses frequently require storage capacities larger than those needed for transmission purposes. One example might be the accumulation of data that is to be uniformly spaced on a magnetic tape arriving from a data source issuing irregularly timed data bits such as a cash register. Another example might be the storage of data for purposes of encoding into error detecting or correcting codes prior to transmission, or the storage of blocks of data for retransmission in case an error is detected.

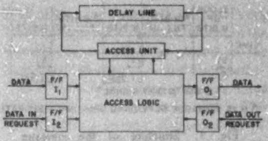


Fig. 5. Buffer store

A versatile buffer store design consisting of a modest amount of circuitry is presented. The storage capacity is larger than the minimum usually necessary for the transmission of data and it is believed that the versatility of the buffer will make it useful for many applications. The buffer uses a delay line looped end to end as its basic storage unit. A magnetostrictive delay line was found to give a large amount of storage capacity and required comparatively small amounts of circuitry.

A block diagram of the buffer is shown in Fig. 5. The message segment in storage is continuously recirculated through the delay line and the several stages of temporary storage (access unit) which are associated with the access logic circuits. The rate of recirculation of the stored segment is fast compared to the shortest interval between successive requests for placing information into the buffer and successive requests for removal of data. Thus, information placed into the buffer has been compressed on a time scale and becomes available for processing at least once between successive input or output operations.

To insert data into the buffer, the data are momentarily stored in F/FI_1. Simultaneously, a synchronous pulse sets F/FI_2, informing the access logic that a bit of data has been stored for transfer into the delay line. The access logic will subsequently recognize the end of the stored message segment when it appears in the access unit, insert the proper data bit, and reset I_1 and I_2.

Similarly, a request for "data out" appearing at the input to F/FO_2 will immediately be satisfied with a bit of information from F/FO_1, previously stored. The access logic will subsequently recognize the beginning of the stored message

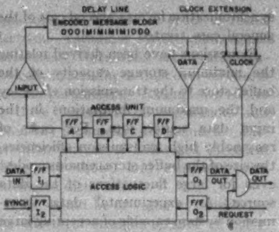


Fig. 6. Magnetostrictive delay line buffer store

segment of information, remove the first data bit, place it in O_1 and reset O_2.

The delay line employed in such a circuit may either be of the digital type such as a shift register or of the analog type such as an electrical or acoustical delay line. The logic and temporary storage elements employed may be of any type meeting sufficient speed requirements.

A magnetostrictive delay line buffer is adaptable to solving the data rate matching problems in data transmission terminals. Such a buffer store is shown in Fig. 6. If, for example, a 200-μsec delay is chosen, a maximum input or output bit rate of 5,000 bits per second can be achieved. Allowing a 2-μsec spacing between pulses, the delay line capacity becomes 100 bits.

Since the amount of data stored in the buffer may vary between zero bits and full, it is necessary to locate both the beginning and the end of the stored message segment. Magnetostrictive delay lines, similar to most other delay lines, perform best when operated in a strictly binary manner. The beginning and end of the stored segment could be located, for example, by reserving specially coded characters to mark the extremities. This, of course, places restrictions on the message. When restrictions on the message are undesirable, a separate delay line may be used to locate the extremities of the stored segment.

Location of the message segment can be advantageously accomplished by encoding the entire stored message segment on the delay line. Such an approach places no restrictions on the incoming message, nor does it require the operation of a separate delay line. An example of such a code is shown in Fig. 6. Fiducial marks in the form of binary "one's" (1) are alternated with message bits (M). The message segment starts and ends with a fiducial "one" marker. Outside of the stored segment are only binary zeros. The access logic can uniquely recognize both the beginning and end of the stored segment when these extremities pass through the four stages of the access unit, and can act upon them accordingly. That is, to insert a bit of information into the delay line, the access logic waits until the access unit is in the state

A B C D
0 0 1 M

(see Fig. 6) and then inserts the message bit into B while A is changed to a marker "one." Similarly, to remove a bit from the delay line, the access logic awaits the pattern

A B C D
M 1 0 0

and then removes the message bit in A while resetting both A and B to zero. The code used on the delay line is 50% redundant. Codes of lower redundancy requiring more complex logic such as

001$MM$1$MM$1$MM$100

or

001$MMM$1$MMM$1$MMM$1000

can easily be constructed.

It should be noted that message bits entering the buffer are always added to the front of the stored segment. To add incoming message bits to the end of the segment would complicate the generation of retiming pulses, which are discussed in the following paragraphs.

Clock pulses are needed to shift the information through the access unit, and also to space new information on the delay line properly. Any error in the timing of the clock circuit, with respect to the pulses on the delay line, could be accumulative. By starting the clock at the time the message segment emerges from the line and by stopping the clock four time intervals after passage of the last stored data bit, the accumulation of any error can be limited.

Generally, such a clock would consist of a complex circuit. The previously described encoding procedure however, permits a clock to be pulled into phase by the marker "one's," thus relaxing the tolerances on the clock.

A very simple self-synchronizing clock circuit may be obtained by physically extending the delay line beyond the data transducer and making use of the marker pulses on the delay line. Four clock transducers are placed on this extension, each an additional bit interval away from the data output transducer; see Fig. 6. The outputs from these four additional transducers are led to a common limiting amplifier. Because of the coding employed on the line, accurately timed clock pulses will be emitted from the common amplifier starting when the first data bit of the message segment appears at the data output transducer and continuing until the last bit has left the access unit.

When the buffer is used for data transmission, as described in previous sections, sufficient logic must be added at the transmitter to permit the buffer to form data into the block structure shown in Fig. 1 in response to a clock operating at precisely R_L bits per second. The time interval between the end and beginning of the circulating message segment stored in the buffer gives a measure of the number of data bits presently in the buffer. When this interval has become sufficiently short a starting signal is generated and N-S bits are counted out of the buffer.

Similarly, at the receiving terminal sufficient logic must be added to allow the buffer to detect the starting signal of a block of incoming data, remove it, and count and store the next N-S bits. An inhibiting signal will be generated when the receiving buffer is empty so that the data receiving machine will be prevented from requesting data when none is available. The data receiving machine may remove data from the receiving buffer at any rate as long as the buffer is not allowed to overflow.

An Experimental Data Transmission System

Two buffer stores of the type described were built for use in an experimental data transmission system. Each buffer uses a 200-μsec magnetostrictive delay line with clock coils spaced to give clock pulses 2.0 μsec apart. Allowing a reasonable interval between the beginning and the ending of the stored message segment the capacity of the line is approximately 90 bits. The coding used on the delay line is 50% redundant leaving a useful capacity of 45 bits. The stored information is circulated 5,000 times per second. Because of precautions taken to prevent any logic "races" the allowable input and output rate is one half the circulation rate, or, in this case, 2.5 kc. If a lower access rate is acceptable, greater storage may be obtained simply by increasing the delay. An experiment has been conducted that inserted an additional 300-μsec delay line in the delay loop. The circuits performed satisfactorily and gave a useful capacity of 118 bits with an access rate of 1 kc.

An experiment was also conducted incorporating the buffer stores into a data transmission system shown in Fig. 7.

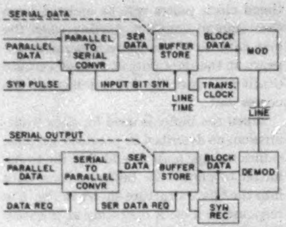


Fig. 7. Data transmission experiment

Data arriving from a data source in a serial manner are put directly into the transmitting buffer together with the "Input Bit Synch" pulse. If the data arrive in parallel form, conversion to serial form is necessary as shown in Fig. 7.

The transmitting buffer forms blocks of *N* bits which are offered for proper modulation at the synchronous transmission rate as determined by the transmission clock.

At the receiver the signal is demodulated. The starting signal is removed from the blocks of *N* bits and the information bits are placed in the buffer. These bits are then ready for delivery to the data receiving unit at the latter's request. Thus, the data receiving unit need not be aware of the nature of the message structure during transmission. Again, either serial or parallel data can be delivered.

The experimental system used a parallel to serial converter of standard design. A frequency shift transmitter and zero axis crossing detector were employed.[3] The transmission clock and bit synch recovery were those which were described in reference 1.

The system could accept input data rates fluctuating arbitrarily between 0 and 975 bits per second. The starting signal consisted of a single bit and the transmission block was 40 bits long. The fixed data transmission speed was 1,000 bits per second.

Conclusions

The many different applications for data transmission pose the question whether a single system can be designed that is capable of serving a large variety of data sources differing widely in their data rate, format, and code. The most severe demands on such a system would occur when the data source rate fluctuates arbitrarily within limits and the data transmission system operates synchronously between transmitter and receiver at a fixed transmission speed. The previous sections have shown that a buffer store may be designed that would permit such a totally asynchronous source to transmit over the synchronous transmission system provided that the average bit rate of the source does not exceed that of the transmission system. The case of a transmission system operating on start-stop bit synchronization forms a special case of the general case treated.

Expressions have been derived relating the minimum storage capacity of the buffer store to the transmission efficiency and the maximum fluctuations in the input data source rate. For cases of reasonably high transmission efficiencies, the size of the buffer store remains modest even for large fluctuations of the data source. An experimental data transmission system capable of accepting large fluctuations in the data input rate was built and tested.

A buffer store using a magnetostrictive delay line looped end to end and capable of performing all the necessary functions was built. It is felt that this design forms a general logic performing element possessing considerable versatility. For example, the looped delay line circuit appears well suited for the instrumentation of encoders and decoders of error detecting or correcting codes.

References

1. Synchronized Clocks for Data Transmission, J. O. Edson, M. A. Flavin, A. D. Perry. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 832–36.

2. An Experimental Modulation-Demodulation Scheme for High Speed Data Transmission, E. Hopner. *IBM Journal of Research Development*, New York, N. Y., vol. 3, Jan. 1959, pp. 74–84.

3. A Frequency-Modulation Digital Subset for Data Transmission Over Telephone Lines, Laurence A. Weber. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 867–72.

Electrical Uniformity and Deviations Control in Polyethylene-Insulated Multipaired Cables

W. L. ROBERTS
ASSOCIATE MEMBER AIEE

THE DEMAND for more and better communications facilities has been an anticipated product of our population and economic growth. Means for extending and improving these facilities in the telephone industry, e.g., nationwide dialing, higher capacity carrier systems, negative-impedance repeaters, data transmission systems, have necessitated adoption of more stringent transmission objectives which affect all levels of service to varying degrees.[1,2]

Cable is a basic element of the telephone system. Improvements in cable uniformity can never guarantee attainment of a desired set of transmission criteria, but will always augment over-all system quality. A substantial over-all improvement requires the concerted reduction of variations contributed by all elements comprising the system. This paper will be restricted to variations in cable parameters, from which effects on a particular system can then be derived.[3,4]

Detailed transmission and crosstalk on paired polyethylene cable have been published recently by Eager, Jachimowicz, Kolodny, and Robinson.[5] Also included are data on variation of parameters with temperature and humidity, and measurements of shield parameters on helical and folded corrugated aluminum shields.

This paper is concerned primarily with variations in primary electrical parameters between pairs within a cable (transverse deviations), and variations between different lengths of cable (longitudinal deviations). Primary parameters are considered in terms of the basic physical dimensions of the cable, and approximate

---

Paper 61-779, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 23, 1961; made available for printing April 25, 1961.

W. L. Roberts is with Superior Cable Corporation, Hickory, N. C.

The author wishes to acknowledge the co-operation of the staff of the Superior Cable Corporation, in particular F. E. Manuel, P. F. Burch, R. N. Kahill, J. D. Sherrill, and D. K. Baker, whose assistance made this study possible.

equations are derived in order to relate electrical variations with dimensional variations. Deviations are treated statistically in arriving at their over-all effects on cable parameters. Examples of effects of these variations on unbalances and secondary parameters are included. Distribution curves of measured electrical values are shown and compared with calculated distributions.

This should be of value in considering the effects of straight splicing of color-coded cable as compared with random splicing. Estimates of the susceptibility of various cable designs to deviation may be made by using approximate expressions derived in the next section. An understanding of some of the sources and types of deviations may be of some benefit in transmission maintenance, especially with plans for upgrading existing cable facilities.

Limitations and other restrictions of approximations are discussed during the development of each equation.

Calculated Deviations of Primary Parameters

The equations for the primary parameters used in this section are approximations, each being exact for some limiting condition. All may be applied to multipaired cable, though some ambiguity may exist concerning the dimensions. The effective shield diameter of a pair is generally the dimension most difficult to determine precisely. The same pair may have a different effective shield diameter depending upon the parameter considered.

An equation in differential form is developed for each of the four parameters. The differential expressions are approximate for two reasons: (1) They are derived from approximate equations, and (2) only first-order effects are considered since only first-degree partial differentials are employed. A third contribution to their approximate nature arises when cable dimensions are considered but, it is believed, all major variables are included in order and the equations are sufficient to allow comparison of the relative contributions from each of the basic dimensions.

MUTUAL CAPACITANCE

Mutual capacitance is the total capacitance appearing across the terminals of a 2-conductor circuit. The total capacitance consists of two components: (1) the direct capacitance between the two conductors, and (2) the capacitance coupling the two conductors through intermediate conductors located within the electric field.

Maupin has published results of an investigation of the effective dielectric constant and geometric parameters of pairs in multipaired cables.[6] These properties were determined by inductance measurements and from formulas developed by Mrs. S. P. Mead for an ideal balanced, shielded pair having the same capacitance and inductance as the cable pair. Mrs. Mead's equation for mutual capacitance[7] is listed in Appendix I.

For purposes of calculating deviations, the familiar approximate expression is sufficient:

$$C/n \doteq \frac{0.01944\epsilon}{\log_{10}\left[(2S/c)\left(\frac{1-(S/D)^2}{1+(S/D)^2}\right)\right]} \text{ mfd (microfarad)/mile} \qquad (1)$$

where ϵ is the effective relative dielectric constant, S is the interaxial pair spacing, c is the conductor diameter, D is the shield diameter, and n is the pair length in miles.

By employing an infinite series in powers of $(S/D)^2$, equation 1 can be reduced to the following:

$$C \doteq \frac{0.01944\epsilon n}{\log_{10}(S/c)+\log_{10}(2)-0.8686(S/D)^2} \text{ mfd} \qquad (2)$$

where $(S/D)^6 \ll 1$.

When this equation is applied to a cable pair, the interaxial spacing is considered equivalent to the average of the insulated conductor diameters (DOD's) and, similarly, c is the average of the diameters of the two conductors. Obviously, this would not be accurate where large differences existed between corresponding conductor dimensions.

If the variations of the independent variables in equation 2 are assumed small, the total differential[8] is useful in determining the resultant effects on the capacitance. The relative (or percentage) change is usually required. Thus, for equation 2

$$\frac{dC}{C} \doteq \frac{\epsilon}{C}\frac{\partial C}{\partial \epsilon}\frac{d\epsilon}{\epsilon} + \frac{S}{C}\frac{\partial C}{\partial S}\frac{dS}{S} + \frac{c}{C}\frac{\partial C}{\partial c}\frac{dc}{c} + \frac{D}{C}\frac{\partial C}{\partial D}\frac{dD}{D} + \frac{n}{C}\frac{\partial C}{\partial n}\frac{dn}{n} \qquad (3)$$

After obtaining the partial derivatives and substituting:

$$\frac{dC}{C} \doteq \frac{d\epsilon}{\epsilon} + \frac{dn}{n} + \frac{22.34C}{\epsilon n} \times \left\{\frac{dc}{c} - [1-4(S/D)^2]\frac{dS}{S} - 4(S/D)^2\frac{dD}{D}\right\} \qquad (4)$$

Difficulty in applying these results to multipaired cable arises in regard to the effective dielectric constant of the nonhomogeneous dielectric and the effective pair-shield diameter. Experience with paper and with polyethylene cables has shown that the effective dielectric constant is linearly associated with the space-per-pair (the proportionate share of the total cable cross section occupied by one pair). On a similar basis it is assumed that the effective D is a function of the cable core diameter. These assumptions are expressed by the two following equations:

$$\epsilon \doteq K - h_1 D_0^2 \qquad (5A)$$

and

$$D \doteq h_2 D_0 \qquad (5B)$$

where D_0 is the cable core diameter, K is a constant, and both h_1 and h_2 depend upon the number of pairs, conductor diameter, stranding configurations of a particular design, and other factors. Under certain considerations h_1 and h_2 may be regarded as independent variables. Equations 5A and 5B may be differentiated and substituted into equation 4, remembering that these are approximations valid only for a narrow range of deviations.

Inspection of equation 4 reveals that the coefficients of the differentials are all of the same order of magnitude for practical values of C (e.g., if $\epsilon = 1.9$, $S/D = 0.36$, and $C/n = 0.083$ mfd, the coefficients range from 0.5 to 1). One further observation applying to multipaired cables involves the coefficient of the dS/S term in equation 4. A variation in S (due to a change in DOD) is accompanied by a corresponding variation in the effective diameter of the shield formed by the remaining pairs. Thus, where S/D is not too different from 0.50, its value is affected only slightly. To account for this, variations in D dependent on S may be assumed.

The capacitance-to-ground component of mutual capacitance is of interest both from the viewpoint of mutual capacitance studies and in the consideration of capacitance unbalance to shield.

By assuming that $(c/S)^2 \ll 1$, the following approximate expression can be derived from the more precise equation for capacitance-to-ground listed in Appendix I of Maupin's paper:[6]

$$C_g/n \doteq \frac{0.03889\epsilon}{\log_{10}\left[\frac{D^4-S^4}{2cSD^3}\right]} \text{ mfd/mile} \qquad (6)$$

Following the same procedure used in deriving equation 4,

$$\frac{dC_g}{C_g} \doteq \frac{d\epsilon}{\epsilon}+\frac{dn}{n}+\frac{11.17C_g}{\epsilon n}\times \left\{\left(1+\frac{4S^4}{D^4-S^4}\right)\frac{dS}{S}+ \frac{dc}{c}-2\left(1-\frac{2D^4}{D^4-S^4}\right)\frac{dD}{D}\right\} \quad (7)$$

For an 0.083-mfd/mile polyethylene-insulated cable, these coefficients range from about 0.55 to 1.1. In applying this expression to unbalance-to-shield computations, it must be noted that the two conductors are assumed equal in diameter and symmetrically disposed with respect to the shield. This does not impose a serious restriction, for small deviations in one conductor do not appreciably influence the capacitance-to-ground of the other conductor by virtue of the assumption $(c/S)^2 \ll 1$.

INDUCTANCE

The low-frequency self-inductance of two parallel wires in free space (or within a nonmagnetic shield) is

$$L_1/n = 1.482 \log_{10}(2S/c)+0.1609 \quad \text{mh (millihenry)/mile} \quad (8)$$

The method employed in the previous section yields

$$\frac{dL_1}{L_1} \doteq \frac{dn}{n}+\frac{0.6436n}{L_1}\frac{dS}{S}-\frac{0.6436n}{L_1}\frac{dc}{c} \quad (9)$$

Equation 8 is a limiting expression which does not take into account proximity effects between the two conductors and between conductors and shield.

At high frequencies, the internal inductance of the conductors and shield is small. The inductance approaches a limiting value determined by the external inductance of the system. For an ideal, balanced pair with homogeneous dielectric, the external inductance[9] is given by:

$$L_e/n = \frac{0.02881n\epsilon}{C} \quad \text{mh/mile} \quad (10)$$

Maupin applied this relation, with a moderate degree of success, to multipaired cable, as an independent check on his method of determining the effective dielectric constant. His results (17% discrepancy in calculated dielectric constant) appear sufficient to warrant use of equation 10 in obtaining an approximation for high-frequency inductance deviations in multipaired cable. (Actually, application of equation 1 implicitly assumes the validity of equation 10 for multipair cable since Mrs. Mead's capacitance deviation employs this equality.)

Proceeding to the total differential,

$$\frac{dL_e}{L_e} \doteq 2\frac{dn}{n}+\frac{d\epsilon}{\epsilon}-\frac{dC}{C} \quad (11)$$

and, substituting from equation 4 and equation 10,

$$\frac{dL_e}{L_e} \doteq \frac{dn}{n}+\frac{0.6436n}{L_e}\times \left\{[1-4(S/D)^2]\frac{dS}{S}-\frac{dc}{c}+4(S/D)^2\frac{dD}{D}\right\} \quad (12)$$

Equations 9 and 11 predict an increase in high-frequency inductance variations as compared with low-frequency values. The terms in equation 11 not common to equation 10 represent variations due to influence of the shield. In applying this to telephone cable, the assumption of equality to the effective shield diameter as employed in the capacitance equation may not be precise.

RESISTANCE

A very thorough compilation of equations and graphs of cable pair resistance as a function of frequency has been provided by Eager, et al.[10] Skin effect, proximity, and eddy current losses are all taken into account. The high-frequency approximations employed in this section are based on these equations.

Because of the complexity of the equations, deviations for low-frequency and high-frequency resistance are considered separately.

The low-frequency d-c resistance is

$$R/n = \frac{0.1095}{\sigma c^2} \quad \text{ohms/mile} \quad (13)$$

where σ is the conductivity relative to copper.

Then

$$\frac{dR}{R} \doteq \frac{dn}{n}-\frac{d\sigma}{\sigma}-2\frac{dc}{c} \quad (14)$$

For high frequencies

$$R_h/n \doteq \frac{1}{3c}\sqrt{\frac{f}{\sigma}} \quad \text{ohms/mile} \quad (15)$$

and

$$\frac{dR_h}{R_h} \doteq \frac{dn}{n}-\frac{1}{2}\frac{d\sigma}{\sigma}-\frac{dc}{c} \quad (16)$$

where f denotes the frequency in kc.

Equations 15 and 16 do not account for proximity effects or eddy current losses in the surrounding shield and pairs; variations due to these effects are estimated in Appendix II. The resulting expression is

$$\frac{dR^1_h}{R_h^1} \leq \frac{dn}{n}-\frac{1}{2}\frac{d\sigma}{\sigma}-\frac{c^2}{S^2-c^2}\frac{dS}{S}- \left[\frac{S^2-2c^2}{S^2-c^2}\right]\frac{dc}{c}-\frac{c\sqrt{S^2-c^2}}{DS}\frac{dD}{D} \quad (17)$$

Comparing equations 16 and 17 with equation 14, the high-frequency resistance is not nearly so sensitive to diameter and conductivity variations as is the case at low frequencies. This advantage is partially offset, however, by new sources of deviations at high frequencies, i.e., variations in proximity and shield losses.

CONDUCTANCE

Conductance in polyethylene multipaired cable is normally so minor in its effect on secondary parameters that it may be neglected throughout the useful frequency range of the cable. Generally, variations in power factor dominate conductance deviations. Even minute traces of contamination in polyethylene affect the power factor noticeably since the initial value is very low (less than 0.0005). Surface conditions on the insulated wire may produce a substantial variation in the effective power factor of a cable pair.

Should it be of interest with other dielectric materials or at high frequencies, conductance may be expressed in terms of capacitance and the effective power factor of the air–dielectric combination:

$$G/n = 6283f(pf)C \quad \text{micromhos/mile} \quad (18)$$

where f is frequency in kc, and (pf) is the power factor. Thus,

$$\frac{dG}{G} \doteq \frac{dn}{n}+\frac{d(pf)}{(pf)}+\frac{dC}{C} \quad (19)$$

Practical Considerations

USE OF STATISTICAL TECHNIQUES

The differential expressions of the preceding section generally cannot be applied directly to a particular pair in a cable. Of those physical dimensions that can be measured, only a few lineal inches at the cable ends are accessible for measurement. Variations occur throughout the length, and a single measurement is expected to differ from the true dimensional average. However, when measurements are made on a large number of pairs, the average and fluctuations about the average yield information concerning the expected electrical variations of the pairs.

Probability theory has been used extensively by the telephone industry in applications such as crosstalk[11] and effects of impedance irregularities,[3,4] to mention only a few.

If the physical dimensions of the pairs (also dielectric constant and conductivity) vary independently and randomly, then each of the primary parameters will also vary randomly, but related through the

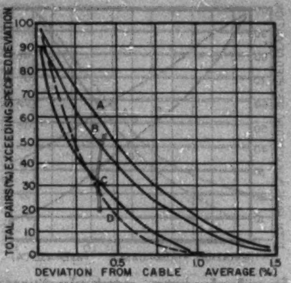


Fig. 1. Deviations in loop resistance of four 2,500-foot lengths, 25-pair 19 Awg (American wire gage) cable

A—Average of individual lengths
B—After color-to-color splicing
C—After random splicing
D—Calculated for random splicing

appropriate functions described in the previous section. If it be further assumed that the dimensional variations occur in accordance with the normal probability distribution, well-known statistical methods may be applied. The techniques developed in the subject of errors of measurement[12] are readily adaptable if the deviations are treated as errors. In practice, relatively small samples are used to estimate deviations and, consequently, slight modifications are demanded by sampling theory. These do not impair the methods described; the modifications are significant only where little is known about the average and standard deviation of the population from which the sample is drawn. It is assumed here that these are already known.

By employing a method known as "compounding of errors,"[13] the standard deviation of any of the primary parameters of the preceding section may be determined by proper use of the appropriate differential approximation. The following equations illustrate the technique:

$$\frac{dy}{y}=a_1\frac{dx_1}{x_1}+a_2\frac{dx_2}{x_2}+a_3\frac{dx_3}{x_3}+\ldots+a_n\frac{dx_n}{x_n} \qquad (20A)$$

$$(s_y/y)^2=(a_1s_1/x_1)^2+(a_2s_2/x_2)^2+(a_3s_3/x_3)^2+\ldots+(a_ns_n/x_n)^2 \qquad (20B)$$

where s_n represents the standard deviation of the nth variable and x_n is its average value.

APPLICATIONS TO PRIMARY PARAMETERS

While application of equation 20B is sufficient to characterize the variations of a particular primary parameter, dimensional distributions are usually based on individual conductors. It will be recalled that average values of the two conductors of the pair were employed in the derivations. The theory of errors again provides a means for coping with this difficulty. The standard deviation of the mean is given by

$$s_m=s/\sqrt{m} \qquad (21)$$

where s is the standard deviation of the population and m is the number of individuals averaged to obtain the mean.

The standard deviation of the average conductor diameter of the pairs is then $(s_c/\sqrt{2})$.

Consider, as an example, equation 14. The standard deviation of the low-frequency loop resistance is

$$s=\frac{R}{\sqrt{2}}\sqrt{(s_n/n)^2+(s_\sigma/\sigma)^2+4(s_c/c)^2} \qquad (22)$$

The standard deviations considered thus far recognize only variations from conductor to conductor within a given length or between different lengths of cable. Variations along the length of a conductor are reflected in the variations in a set of dimensional measurements, made on short samples. Distinction between the two bases of calculation is important for their standard deviations may be quite different. Both contribute to variations of primary parameters but to a different degree depending upon the length of the variation "cycle" compared with the conductor or pair length.

Each conductor (and pair) of a cable has been processed through several distinct stages in manufacturing, e.g., wire drawing, wire coating, twinning, stranding, etc. At each stage are independent sources of variations. Each source may be random in nature but specific electrical properties of the cable may be essentially independent of certain of these sources.

The design of the cable may contribute sources of deviations which are not random in nature. The lengths of different pairs in a cable are not all equal; different twist lengths are used for crosstalk reduction. Also, differences exist because the length of the helical path in an outside layer differs from that in an inside layer. In the previous example on loop resistance, these effects appear in the variation of length term (s_n/n), but the two conductors of a pair are affected equally, not independently as was assumed.

Equation 21 is suitable for use in comparing deviations in spliced cable lengths. Since the standard deviations of the primary parameters are proportional to the corresponding average values, it is convenient to work with the coefficient of variation, ν. Thus, for m lengths, the coefficient of variation of pair loop resistance should decrease in accordance with the following:

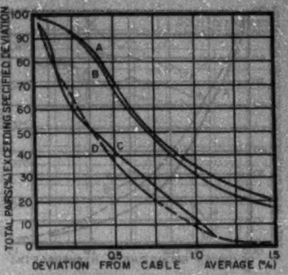


Fig. 2. Deviations in mutual capacitance of five 4,000-foot lengths, 18-pair 19 Awg cable

See subcaption, Fig. 1

$$\nu_m=\frac{\nu}{\sqrt{m}} \qquad (23)$$

Fig. 1 illustrates this result graphically. The distribution showing pair resistance of the single lengths is based on four groups of 25 pairs. The two 50-pair cables were consecutive production lengths. In the graph, results are shown for random and color-to-color splicing techniques. The results from random splicing are in fair agreement with the calculated distribution but the color-spliced values are not too different from the original distribution. Fig. 2 shows results of a similar study employing five consecutively manufactured lengths of 18-pair cables.

The grouping from sequential production lengths should indicate an extreme condition for results from color-to-color splicing. Those nonrandom departures which persist from length to length may deteriorate return losses on loaded, repeatered cable. However, they could contribute toward a slight improvement in loaded cable except for deviations in cutoff frequency.

UNBALANCES AND SECONDARY PARAMETERS

The approximations for deviations of the primary parameters may be conveniently employed for analysis of resistance and capacitance unbalance. An example of this use is offered for use in capacitance unbalance-to-shield estimates. This unbalance is the difference in the direct capacitance-to-shield of the two conductors of a pair and, for a particular construction, should be proportional to capacitance-to-ground.

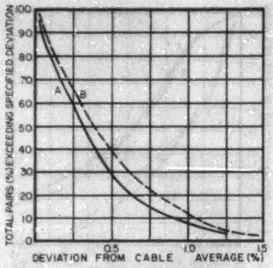


Fig. 3. Deviations in 1-kc attenuation

A—Measured
B—Calculated from resistance and capacitance deviations

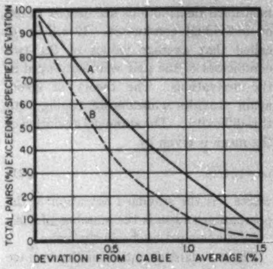


Fig. 4. Deviations in 1-kc characteristic impedance

See subcaption, Fig. 3

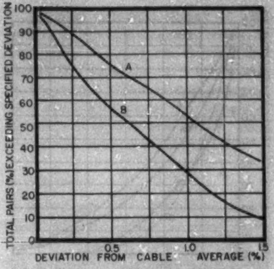


Fig. 5. Transverse mutual capacitance deviations, 50-pair 19 Awg aerial cable

A—Before tolerance and design changes
B—After revision

Then,

$$U_C/n \doteq q(C_{g1}-C_{g2}) \quad \text{mmf (micromicrofarads)/mile} \qquad (24)$$

where q is a constant, U_C is capacitance unbalance-to-shield, and C_g is taken from equation 6. But, U_C is itself a deviation since $C_{g1} \doteq C_{g2}$. Consequently, the expression can be interpreted as

$$dU_C/n \doteq q(dC_{g1}-dC_{g2}) \text{ mmf/mile} \qquad (25)$$

Since both C_{g1} and C_{g2} are taken from the same population, the standard deviation is

$$s_{U_0}/n \doteq q\sqrt{2}(s_{C_g}) \text{ mmf/mile} \qquad (26)$$

and s_{C_g} is obtained from equation 7.

This is equivalent to the rms unbalance because the average (algebraic) should equal zero. The average unbalance as normally defined is actually the mean deviation and is related to the rms unbalance by

$$\text{Avg}\,(U_C/n) \doteq 0.7979\,(s_{U_0}/n) \doteq 0.8q\sqrt{2}(s_{C_g}) \quad \text{mmf/mile} \qquad (27)$$

The value of q will range from about 0.2 to 0.6 for pairs in an outside layer of a cable, the lower values applying to unit constructions where some shielding is offered by neighboring units.

Consistent deviations in individual conductor values of capacitance-to-ground result from helix length variations between different pairs but, in contrast to the effect on primary parameters, this type of variation does not contribute to unbalances in general.

Secondary parameter deviations can be determined by use of the fundamental equations written in terms of the primary parameters. Hebbert derived an equation of this type for attenuation.[14] An expression for deviations in the absolute value of the characteristic impedance is included in Appendix III of this paper. Substitution of applicable equations written in terms of dimensions reveals sources of deviations which are not otherwise immediately obvious. Consider the approximate expression for attenuation at 1 kc:

$$\alpha/n \doteq 0.488\sqrt{RC} \text{ decibels} \qquad (28)$$

and

$$\frac{d\alpha}{\alpha} \doteq \frac{1}{2}\frac{dR}{R}+\frac{1}{2}\frac{dC}{C} \qquad (29)$$

Comparing equations 4 and 11, it may be observed that variations in conductor diameter will simultaneously affect both capacitance and resistance, and will produce a slight compensating effect on attenuation. Increasing the conductor diameter will cause an increase in capacitance and a decrease in resistance. Consequently, variations in attenuation should be slightly better than would be expected from the individual distributions of pair resistance and capacitance. These differences are illustrated in Fig. 3.

These effects do not always result in improvements. The corresponding characteristic impedance approximation at 1 kc is

$$Z_0 \doteq 12.6\sqrt{\frac{R}{C}}\ \angle\,45^\circ \qquad (30)$$

and

$$\frac{dZ}{Z} \doteq \frac{1}{2}\frac{dR}{R}-\frac{1}{2}\frac{dC}{C} \qquad (31)$$

In this case conductor diameter variations tend to increase impedance variations above the expected value. Compare Fig. 4 for the same cable used for attenuation variations. A similar effect also occurs at high frequencies where inductance and capacitance dominate the magnitude of the characteristic impedance.

The examples cited here are sufficient to illustrate some of the ways in which the differential expressions for the primary parameters can be employed. It is believed that they might be of value as a guide in predicting performance of various cable designs if only from a comparative viewpoint based on known performance characteristics of an existing cable design. Both primary and secondary parameters, including unbalances, generally become less susceptible to variation as the nominal capacitance decreases.

APPLICATIONS TO MANUFACTURING CONTROL DEVIATIONS

The many factors affecting electrical properties of a telephone cable are compounded by the various stages in the manufacturing process. The results of a set of electrical measurements cannot usually be translated directly into specific sources of particular deviations appearing in the finished product. However, statistical techniques provide a powerful tool for analysis, especially when it is possible to convert the results into their fundamental equivalents of dimensions, dielectric constant, and conductivity.

Variations and combinations of the equations listed in this paper have been successfully applied, using only simple statistical relations and a great deal of the knowledge and observations of personnel associated with the manufacturing process.

An assumption was made in the previous section that all deviations followed the laws of the normal probability distribution. The realization of this assumption is the exception, rather than the rule. But the departures from normality usually provide the clues of most benefit in improving design and manufacturing techniques.

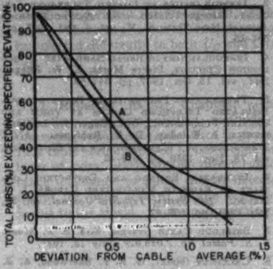


Fig. 6. Transverse mutual capacitance deviations, 7-quad 19 Awg cable

See subcaption, Fig. 5

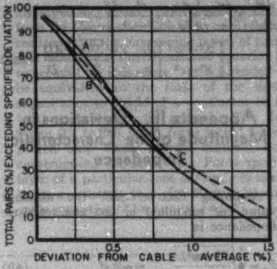


Fig. 7. Loop resistance deviations

A—Measured
B—Measured and corrected for design variations
C—Calculated from dimensional deviations

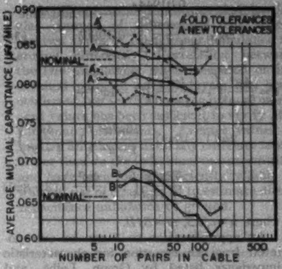


Fig. 8. Range in average mutual capacitance (95% of reels) versus number of pairs

A—19 Awg aerial exchange cable
B—19 Awg toll grade aerial cable

Figs. 5 and 6 show results obtained from an exchange-type aerial cable and a quadded-type toll cable before and after making design and dimensional tolerance changes. The design of these two particular cables has been more or less optimized; naturally the same degree of uniformity cannot be achieved simultaneously on all types of a particular series of cables because certain compromises sometimes must be made in specific designs. There are instances where the addition of one or two more pairs in a particular cable can result in a substantial improvement in uniformity. Restrictions in geometric configurations of the pairs in a cable are generally the most severe in the smaller pair-count types.

Fig. 7 contains a graph of a calculated pair resistance distribution based on equations of the previous sections and compares this with a distribution of measured values. The third distribution in Fig. 7 has been corrected for calculable deviations due to twinning and stranding differences.

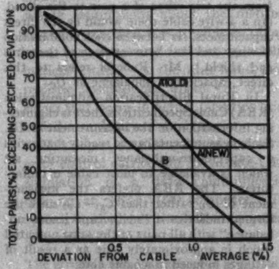


Fig. 9. Transverse mutual capacitance deviations (95% of reels)

A—19 Awg aerial exchange cable (18 through 1,000 pairs)
B—19 Awg toll grade aerial cable

The remaining graphs show variation ranges and distributions of capacitance, inductance, and pair resistance in both aerial-exchange-type and paired toll grade cables. Figs. 8 and 9 show, for comparison, limits of capacitance deviations before initiation of changes deemed desirable as a result of design and dimensional analyses. The values shown are representative of normal production and it is believed that further application of the methods described will result in additional improvements.

Conclusions

Approximations of expected variations of primary electrical parameters suitable for use with multipaired telephone cable have been developed in terms of physical dimensions, effective dielectric constant, and conductivity. Simple statistical techniques, borrowed directly from the subject of errors of measurement, have been applied to several primary and secondary parameters and electrical unbalances. Test data have been presented which substantiate general effects predicted by these approximations.

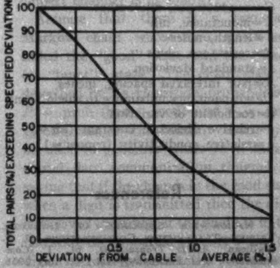


Fig. 10. Transverse inductance deviations (95% of reels), 19 Awg aerial exchange cable (18 through 100 pairs)

Tolerances and design criteria based on these relations have been applied to the control of deviations in telephone cable. Significant improvements in deviations have been noted and continuing improvements are being made to obtain even better control over cable characteristics which are significant in cable applications in the operating telephone industry.

Appendix I. Mutual Capacitance of an Ideal, Balanced Pair

Mrs. Mead's equation for mutual capacitance is

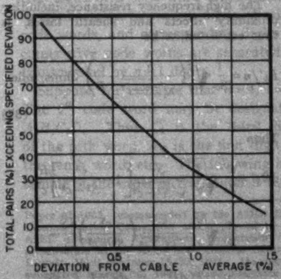


Fig. 11. Direct-current resistance deviations (95% of reels), 19 Awg aerial exchange cable (18 through 100 pairs)

$$C/n=\frac{0.01944\epsilon}{\log_{10}\left[\frac{1}{u}\frac{1-w^2}{1+w^2}\right]-0.1086\delta}\text{ mfd/mile} \tag{32}$$

$$\delta=\frac{4u^2}{1-u^2}\left(1-\frac{4w^2}{1-w^4}\right)(1-4w^2)+2u^4(1+16u^4)-\frac{16u^2w^6}{1-u^2}\left(1-\frac{4w^2}{1-w^4}\right)+\frac{32u^4w^2}{1-w^2}\left(1-\frac{4w^2}{1-w^4}\right)(3-8w^2)$$

where $u=c/2S$, $w=S/D$.

An approximation to δ can be extrapolated from an equation for characteristic impedance listed by Green, Leibe, and Curtis (equation 35 of their paper):[15]

$$\delta\doteq 4u^2(1+u^2)(1-4w^2) \tag{33}$$

When $w<0.5$ and $u<0.3$, the capacitance will be correct to better than 0.5% if the following approximation is employed:

$$\delta\doteq 4u^2(1+u^2)(1-4w^2)^2(1+40u^2w^2) \tag{34}$$

Appendix II. Variations in High-Frequency Resistance

Using equations and graphs listed by Eager, et al.,[10] variations due to proximity and shield losses may be estimated from equations credited to A. H. M. Arnold and L. Meyerhoff, respectively.

$$P=\sqrt{\frac{1}{1-(c/s)^2a(x)}}=\text{proximity factor} \tag{35}$$

The high-frequency resistance equation 15 is multiplied by P. The value of $a(x)$ is graphed in Fig. 24 of Eager's paper, but varies slowly at high frequencies and is taken as unity for simplification of the estimate for high-frequency resistance variations.

Meyerhoff's formula for sheath losses at high frequencies can be reduced to the following approximation:

$$\Delta R/n\doteq\frac{1}{3D}\sqrt{\frac{f}{\sigma}}\text{ ohms/mile} \tag{36}$$

where ΔR is the increase in pair resistance.

The high-frequency resistance, including proximity effects and sheath losses, is thus estimated to be

$$R_h'/n\doteq\frac{1}{3c}\sqrt{\frac{f}{\sigma}}\left(\frac{S}{\sqrt{S^2-c^2}}+\frac{c}{D}\right)\text{ ohms/mile} \tag{37}$$

Then

$$dR_h'\doteq\frac{n}{3c}\sqrt{\frac{f}{\sigma}}\frac{S}{\sqrt{S^2-c^2}}\left[1-\frac{c^2}{S^2-c^2}\right]\frac{dc}{c}-\frac{n}{3c}\sqrt{\frac{f}{\sigma}}\frac{S}{\sqrt{S^2-c^2}}\left(\frac{c^2}{S^2-c^2}\right)\frac{dS}{S}-\frac{n}{3D}\sqrt{\frac{f}{\sigma}}\frac{dD}{D}+R_h'\frac{dn}{n}-\frac{1}{2}R_h'\frac{d\sigma}{\sigma} \tag{38}$$

or

$$\frac{dR_h'}{R_h'}\leq\frac{dn}{n}-\frac{1}{2}\frac{d\sigma}{\sigma}-\frac{c^2}{S^2-c^2}\frac{dS}{S}-\left[\frac{S^2-2c^2}{S^2-c^2}\right]\frac{dc}{c}-\frac{c\sqrt{S^2-c^2}}{DS}\frac{dD}{D} \tag{39}$$

Appendix III. Deviations in Magnitude of the Characteristic Impedance

Assuming that $G\ll wC$, the absolute value (or modulus) of the characteristic impedance is

$$Z_0=\sqrt[4]{\frac{R^2+w^2L^2}{w^2C^2}}\qquad w=2\pi f \tag{40}$$

and

$$\frac{dZ_0}{Z_0}\doteq\frac{R^2}{2(R^2+w^2L^2)}\frac{dR}{R}+\frac{w^2L^2}{2(R^2+w^2L^2)}\frac{dL}{L}-\frac{1}{2}\frac{dC}{C} \tag{41}$$

When $R\ll wL$, substitution of equations 4 and 13 yields:

$$\frac{dZ_0}{Z_0}\doteq-\frac{1}{2}\frac{d\sigma}{\sigma}-\frac{1}{2}\frac{d\epsilon}{\epsilon}-\left(1+11.17\frac{c}{\epsilon n}\right)\frac{dc}{c}+11.17\frac{c}{\epsilon n}\left[\left(1-4\frac{S^2}{D^2}\right)\frac{dS}{S}+4\frac{S^2}{D^2}\frac{dD}{D}\right] \tag{41A}$$

When $R\ll wL$, substitutions are made from equations 4 and 11:

$$\frac{dZ_0}{Z_0}\doteq\frac{22.34C}{\epsilon n}\times\left[\left(1-4\frac{S^2}{D^2}\right)\frac{dS}{S}-\frac{dc}{c}+4\frac{S^2}{D^2}\frac{dD}{D}\right]-\frac{1}{2}\frac{d\epsilon}{\epsilon} \tag{41B}$$

Appendix IV. Nomenclature

c = conductor diameter, inches
C = mutual capacitance, mfd
D = effective shield diameter, inches
D_0 = cable core diameter, inches
$\overline{DOD}$ = diameter-over-dielectric, inches
f = frequency, kc
G = conductance, micromhos
L = inductance, mh
n = length, miles
R = resistance, ohms
s = standard deviation
S = pair interaxial spacing, inches
U_C = capacitance unbalance to shield, mmf
v = coefficient of variation
ϵ = relative dielectric constant (air = 1)
σ = relative conductivity (copper = 1)

References

1. Nation-wide Switching of Intertoll Trunks—Transmission Considerations, O. F. Wallman. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, May 1958, pp. 223–30.

2. Study of Return Losses on Loaded Trunk Cables and Methods Taken to Improve Them, G. H. Speake. *Ibid.*, pp. 201–04.

3. Irregularities in Loaded Telephone Circuits, George Crisson. *Bell System Technical Journal*, New York, N. Y., vol. 4, Oct. 1925, pp. 561–85.

4. Irregularities in Broad-Band Wire Transmission Circuits, Pierre Mertz, K. W. Pfleger. *Ibid.*, vol. 16, Oct. 1937, pp. 541–59.

5. Transmission Properties of Polyethylene-Insulated Telephone Cables at Voice and Carrier Frequencies, G. S. Eager, Jr., L. Jachimowicz, I. Kolodny, D. E. Robinson. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, Nov. 1959, pp. 618–40.

6. Interaxial Spacing and Dielectric Constant of Pairs in Multipaired Cables, J. T. Maupin. *Bell System Technical Journal*, vol. 30, July 1951, pp. 652–67.

7. Shielded Cable System, Mrs. S. P. Mead. *U. S. Patent No. 2,086,629*, July 13, 1937.

8. Calculus (book), G. E. F. Sherwood, A. E. Taylor. Prentice-Hall, Inc., Englewood Cliffs, N. J., third edition, 1954, p. 460.

9. Maupin, *op. cit.*, p. 663.

10. Eager, et al., *op. cit.*, pp. 635–36.

11. Open-Wire Crosstalk, A. G. Chapman. *Bell System Technical Journal*, vol. 13, Jan. 1934, pp. 19–37; Apr. 1934, pp. 195–238.

12. An Introduction to Scientific Research (book), E. B. Wilson, Jr. McGraw-Hill Book Company, Inc., New York, N. Y., 1952, pp. 232–76.

13. *Ibid.*, p. 272.

14. Characteristics of Toll Telephone Cables, C. M. Hebbert. *Bell System Technical Journal*, vol. 20, July 1941, pp. 293–330.

15. The Proportioning of Shielded Circuits for Minimum High-Frequency Attenuation, E. I. Green, F. A. Leibe, H. E. Curtis. *Ibid.*, vol. 15, Apr. 1936, pp. 248–83.

◆

Discussion

A. G. Chapman (Rural Electrification Administration, Washington, D.C.): I would like to comment on equation 24 of Mr. Roberts' paper. There are two methods of measuring capacitance unbalance to ground in a length of paired cable. The measuring frequency is usually 1,000 cycles per second.

The first method is called the "total" capacitance unbalance to ground, and involves grounding all the pairs to the shield except the test pair. This is indicated as $C_{g1}-C_{g2}$ by Mr. Roberts in equation 24. The second method involves measuring the difference of the two direct capacitances from the two wires of a pair to the shield. (In a 4-wire cable there would be ten direct capacitances, six between each combination of wires and four between each of the wires and shield.) Mr. Roberts refers to this direct capacitance unbalance as U_C.

In Rural Electrification Administration (REA) Cable Specifications there is engineering information on the measurement of U_C. Such measurements can be readily made with a capacitance-unbalance measuring set which is available to most cable manufacturers. The REA prefers the measurement of U_C rather than $C_{g1}-C_{g2}$ since the former indicates a "noise capacitance unbalance" with all pairs at the same potential which is approximately true as a result of voltages induced in a cable core.

Mr. Roberts points out that there is a rough correlation q between U_C and $C_{g1}-C_{g2}$ but that q varies from 0.2 to 0.6. The unbalances to ground may exist because (1) the two insulated conductors may vary transversely and (2) the first half of a twist may be longer than the second half, or vice versa. These two causes of unbalances may be reflected differently in both U_C and $C_{g1}-C_{g2}$.

W. L. Roberts: As Mr. Chapman points out, there are two distinct methods of measuring capacitance unbalance to ground. In a cable core composed of more than one layer of pairs, practically no direct capacitance unbalance to shield can be detected except for pairs located in the layer adjacent to the shield. The small unbalances displayed by "inside" pairs generally result from the electrostatic shielding or screening realized from the intervening layers of pairs. The factor q was introduced into equation 24 to account for this effect. It is intended to be equivalent to the ratio of the direct capacitance to shield of a conductor to its total capacitance to ground. The full range 0.2 to 0.6 listed for q would be expected only where several different core constructions are involved. For a specific layer of a particular core lay-up, q is taken to be a constant (e.g., for 50-pair 19 Awg cable consisting of four units, $q \doteq 0.2$).

Mr. Chapman also indicates that particular sources of asymmetry may affect the two types of capacitance unbalance to ground differently. Correlation between the two unbalances when measured on individual pairs in outside layers has been confirmed, but scatter between pairs of measured values has been sufficient to verify Mr. Chapman's comments. Average values of the two unbalances naturally correlate much better than do the individual measurements, and have been found sufficiently reliable to warrant use of equation 24 to estimate direct capacitance unbalance to ground.

Computer Simulation of the Use of Group Codes With Retransmission on a Gilbert Burst Channel

W. R. COWELL
NONMEMBER AIEE

H. O. BURTON
NONMEMBER AIEE

Summary: A study of error control by coding was made by Monte-Carlo simulation of a burst-noise channel on an IBM (International Business Machines) *7090*. Using short group codes, comparisons were made between correction and detection with retransmission. Also, the effect of interleaving the code words was studied. Error detection with retransmission showed a consistently better performance than error correction and the time division resulting from interleaving was effective in combating burst type error patterns.

THE PROBLEM of predicting the performance of error-correcting and error-detecting codes in digital data transmission has been approached from several standpoints.[1-3] If one makes rather simple assumptions about the distribution of errors then it is possible to obtain some very deep and interesting mathematical results. On the other hand, more realistic assumptions about the sort of errors that occur in actual facilities bring added complexity of analysis.

This paper is a discussion of a study of error control carried out by Monte-Carlo simulation of a data communication link on an IBM *7090* computer. The system studied has the following general properties:

The errors in the channel occur in bursts.

The message is encoded into the words of a fairly short (word length ≤ 12) group code (systematic code, parity check code). At the receiver the code is used to correct specified errors and to request retransmission when certain other specified errors are detected. The reverse channel over which such requests are made is assumed to operate without error.

The code words may be transmitted in an interleaved pattern so that between digits of a given word there are transmitted digits of other words.

Let us expand on these assumptions and explain what models were used and how they were incorporated into the computer program.

The Gilbert Burst-Noise Channel

E. N. Gilbert has proposed a model of a channel in which errors occur in bursts.[4] He has shown that his model fits certain telephone plant observations rather well. The reader is referred to Gilbert's paper for details but a brief summary of the Gilbert channel is as follows:

Assume that there is a two-state Markov chain whose two states are called the "good" state and the "bad" (or "burst") state. The chain may be described by its transition probabilities, P = prob (bad | good) and π = prob (good | bad). The Markov chain may be used to generate errors in a binary digital data communication channel by saying that if the chain is in the good state when a digit is transmitted then the digit is received correctly; if the chain is in the bad state the probability is only h that the digit is received correctly. Moreover, the Markov chain is stepped forward (and may or may not change state) after the transmisson of each digit.

Gilbert's model has the virtue of simplicity in that only three parameters (P, π, and h) are needed for a complete description of the channel, and yet the feature of errors in bursts (namely when the chain is in the bad state) is included. In Gilbert[4] the capacity of such a channel is calculated and a method of estimating the parameters from experimental data is described. It is this model that has been employed in simulating a data channel.

Interleaving

If errors occur in bursts, codes may be designed that are particularly suited to detecting or correcting this type of error pattern or the code words may be transmitted in an interleaved fashion so that digits of a code word are separated by digits of other code words. Thus, the effect of a burst is spread over several words and is less likely to exceed the error detecting or correcting capabilities of any one word.

The interleaving plan that has been used in the present study can be described as follows: Let the code words be $n \geq 2$ digits long. Let $s \geq 1$ be an integer. It shall be assumed that the code is interleaved at degree s if any pair of consecutive digits of a code word are separated in transmission by $ns-2$ digits of other code words, and corresponding digits of consecutive code words are separated in transmission by $n-1$ digits.

Suppose, for example that $n=4$, $s=2$. The i^{th} code word may be designated $A_iB_iC_iD_i$. Thus C_5 is the third digit of the fifth word, A_{10} is the first digit of the tenth word, etc. The following is a sample of the order of transmission when

---

Paper 61-1023, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted March 13, 1961; made available for printing August 25, 1961.

W. R. COWELL and H. O. BURTON are both with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

the code is interleaved at degree 2 (transmission takes place from left to right):

$$\dots D_{-6}C_{-4}B_{-2}\underset{\uparrow}{A_0}D_{-5}C_{-3}B_{-1}\overset{\downarrow}{A_1}D_{-4}C_{-2}\underset{\uparrow}{B_0}A_2$$

$$D_{-3}C_{-1}\overset{\downarrow}{B_1}A_3D_{-2}\underset{\uparrow}{C_0}B_2A_4D_{-1}\text{-}$$

$$\overset{\downarrow}{C_1}B_3A_5\underset{\uparrow}{D_0}C_2B_4A_6\overset{\downarrow}{D_1}C_3B_5A_7\dots$$

Since numbering words may be started anywhere the negative subscripts are perfectly acceptable per se. However, suppose the first word of the message is numbered 0. Then those digits with negative subscripts are not message digits and so the negatively indexed digits which follow A_0 may be regarded as dummy digits which are transmitted simply to establish the interleaved pattern.

The arrows below the line point to the digits of word 0. Note that there are $ns-2=6$ digits between each arrow. The arrows above the line point to the digits of word 1. Note that word 1 can be brought opposite the lower arrows by fixing those arrows and translating the entire sequence 4 digits to the left. If a code of length n is interleaved at degree s let the sequence of transmitted digits be observed through n windows so that one digit appears in each window and the interval between consecutive windows is $ns-2$ digit positions. Each code word will appear in the windows as the sequence of transmitted digits is stepped forward. If word i appears in the windows then word $i+1$ will appear in the windows after the entire sequence has been stepped forward exactly n digit positions.

This interleaving technique is treated mathematically in the Appendix where it is demonstrated that interleaving at any degree $s \geq 1$ for any code of length $n \geq 2$ is possible in the sense that each digit position in the transmission sequence contains exactly one digit of a message word or exactly one digit of a start pattern corresponding to the negatively indexed positions in the example. Conversely, there is a digit position in the transmission sequence for every message digit.

Ordinary digit by digit, word by word transmission shall be referred to as "interleaving at degree 0" even though it is not a special case of the foregoing.

The Decoder

In the systems that have been simulated, the redundancy introduced by the encoding process may be used for error correction or error detection with retransmission or some balance between the two. Since the code words of length n form a subgroup of the group of all binary sequences of length n under componentwise modulo 2 addition, it is convenient to describe the operation of the decoder in terms of elementary group theory.

Let the group of all sequences of length n be decomposed into cosets relative to the subgroup of code words. In each coset a coset leader b is selected and each element of the coset is expressed in the form $b+x$ where x is a code word. This expression is unique and it will be assumed that the 0 word is the leader of the coset consisting of the code words. Now divide the cosets into two classes, a "correction" class (which contains at least the coset of code words) and a "retransmission" class. If a received word lies in a correction coset and is of the form $b+x$, it is decoded to x. If a received word lies in a retransmission coset the transmitter is informed via an error-free reverse channel to retransmit the code word. Retransmissions are assumed to occur until a word in a correction coset is received.

Suppose an error word of length n is defined as a sequence of n binary digits in which 1 represents an error digit and 0 represents a correct digit. A Gilbert channel may be thought of in the following terms: the transmitter supplies a code word and the previously defined Markov process supplies an error word. The received word is the componentwise modulo 2 sum of these two words. The aforementioned decoder corrects exactly those error words which are coset leaders of correction cosets and calls for retransmission when the error word lies in a retransmission coset. From the group property it follows that whether a received word causes retransmission, is decoded correctly, or is decoded incorrectly depends only on the error word and not on which code word was transmitted.

As an example we take a code with $n=5$ which has four code words, namely 00000, 01011, 10101, and 11110. If we decide to correct all single errors per word and retransmit when other errors are detected then the first 6 rows in the following array are correction cosets and the last 2 are retransmission cosets:

00000	01011	10101	11110
10000	11011	00101	01110
01000	00011	11101	10110
00100	01111	10001	11010
00010	01001	10111	11100
00001	01010	10100	11111
11000	10011	01101	00110
10010	11001	00111	01100

The Parameters

The parameters that describe the channel, code, and results of the simulation will be discussed in this section.

The channels that have been used are all Gilbert burst-noise channels and each of these is completely described by the three previously defined parameters, P, π, and h.

The codes are described by giving the length of the code words, the number of information places, and the degree of interleaving. Only one code of each length and redundancy was used and so there is no chance for ambiguity. Each code that was used is a "best" group code in the sense of Slepian[3] although we do not claim that it is necessarily the best code when used as a corrector-detector on a burst-noise channel. No permutations for these codes were investigated and their exact parity check rules may be obtained from reference 3 except that the 4,3 code and 3,2 code are simple parity check codes.

Suppose now that a Gilbert channel, a group code together with its interleaving, and a decoder of the type described have been specified. Let code words be transmitted over the channel and imagine an observer who is able to determine and keep a record of the error word associated with each transmitted code word.

The observer classifies the error words into the following three categories:

1. The error words that permit the transmitted word to be decoded correctly at the receiver.
2. The errors words that caused the transmitted word to be incorrectly decoded at the receiver.
3. The error words that cause retransmission of the transmitted word.

In examining the error words we are not distinguishing between the transmissions where the message word is being sent for the first time and the transmissions in which a message word is being retransmitted. As mentioned earlier, the classification of the error words into categories depends only on the error words and is independent of which message word was involved.

Let C be the number of error words in category 1, M the number of error words in category 2, and R the number of error words in category 3.

If F is the probability that a randomly selected message word can be ultimately decoded incorrectly, possibly after retransmissions have occurred, then F may be estimated by

$$\hat{F}=\frac{M}{M+C}$$

If θ is the probability that a randomly selected transmission of a code word will be followed by a retransmission of the same code word, (the selected transmission may be a retransmission, of course) then θ may be estimated by

$$\hat{\theta}=\frac{R}{M+C+R}$$

The efficiency, e, of a group code of length n used with a corrector-detector decoder shall be defined as the ratio of the number, m, of information digits in a code word to the expected number of digits transmitted until a randomly selected message word is decoded at the receiver. Thus, e may be estimated by

$$\hat{e}=\frac{m(M+C)}{n(M+C+R)+LR}$$

where L is the number of digits "lost" or "wasted" when retransmission occurs. These digits will be counted as "transmitted" digits in determining the efficiency even though they may not actually appear on the channel. In a real communication system these lost digits might be digits that would have been sent during the time required to reverse the direction of transmission or they might be digits thrown away because of an interleaved transmission pattern or used as a starting sequence in an interleaved transmission pattern. In determining L the digits of the retransmitted code word itself will not be counted. It is clearly necessary to make some assumptions about L in a simulation program. It was assumed that when a received word was examined by the decoder and found to be in a retransmission coset the lost digits consisted exactly of the $n-1$ digits preceding the first digit of the word under examination, the $n-1$ digits following the last digit of that word, and all the digits between digits of the examined word. These are lost in the sense that transmission equivalent to these digits plus retransmission of the code word itself must occur before the new received word can be examined by the decoder. In the example of the section on interleaving, if the decoder discovered that word 1 must be retransmitted then the lost digits are D_{-5}, C_{-3}, B_{-1}, C_3, B_5, and A_7, as well as all digits between digits of word 1. It is generally seen, under these assumptions, that the following is true:

$$L=2(n-1)+(n-1)(ns-2)=ns(n-1)$$

The redundancy of the code is defined to be $r=1-e$ and the redundancy which is added by retransmission is exactly $r-\frac{n-m}{n}$.

The Computer Program

The computer program generates sequences of ones and zeros by using the previously described Gilbert model. One represents an error and a zero represents no error. The sequences of zeros and ones are grouped into blocks of length n where n is the length of the code being tested. The blocks, which are the error words, are then analyzed by the computer program to determine what effect they would have had on transmitted words. As mentioned in the section on the decoder, only the error words need be examined to determine completely the number of errors, retransmissions, and correct receptions. The program is capable of making the following decisions with respect to the error words:

1. The error word consists entirely of zeros. This implies that the transmitted word is received unchanged.
2. The error word belongs to a correction coset (different from the coset of code words) and is a coset leader. This implies that the received word is corrected and accepted and the accepted word is the same as the transmitted word.
3. The error word belongs to a correction coset (different from the coset of code words) but is not a coset leader. This implies that the received word undergoes a correction process and is accepted but in this case the accepted word is not the same as the transmitted word.
4. The error word belongs to the subgroup of code words but is not the all zero element. This implies that the received word is accepted but is different from the transmitted word.
5. The error word belongs to a retransmission coset. This implies that a retransmission is requested.

Thus, by keeping a record of how many error words fall into each of these categories, the program is able to determine the number of correct receptions, the number of corrections, the number of errors, and the number of retransmissions that would have occurred.

The program can handle codes that are used for correction only, for retransmission only, or for some specified combination of the two. For correction only, decisions 1, 2, 3, and 4 are used. For retransmission only, decisions 1, 4, and 5 are used. All five decisions are used when correction and retransmission are combined.

Production of the actual error words is accomplished by generating pseudorandom numbers in the computer and comparing such numbers with fixed numbers derived from the channel parameters. The program makes use of one of the common ways of generating such numbers, namely the sequence r_1, r_2, . . . defined by $r_1=5^{15}$, $r_{i+1}=$least significant 35 binary digits of $5^{15}\cdot r_i$.

Any group code up to 35 digits in length may be tested. The code words are calculated by the program using a parity check matrix which is read in as data for each run. Codes may be interleaved with the restriction that $(n-1)s\leq 100$.

The output of the program consists of the number of error words which fall into each of the five categories and the relative frequencies of 1, a word accepted in error, 2, a transmission error, and 3, a retransmission. The estimated efficiency and the redundancy added by retransmission are also calculated.

The program was written so it would stop when a certain number of error words fall into category 1 (corresponding to transmitted words unchanged). This number has been 50,000 for all runs described in this paper unless otherwise stated. This resulted in a total sample of 50,000 to 60,000 words except when the probability of error (and hence of retransmission) was very high.

The program was written for the IBM *7090* computer at Bell Telephone Laboratories, Murray Hill, N. J., using the Fortan Assembly Program. The primary method of checking was to use a binary symmetric channel which is a special case of a Gilbert channel. Theoretical results were calculated for several codes and compared with the results obtained by using the program.

The Results

The primary concern has been to try to give a reasonable answer to the question of whether interleaving and retransmission could significantly improve the capabilities of short codes in reducing the error rate in a burst-noise channel. Consequently the program has been used to examine the effect of interleaving and to compare retransmission with correction. Once a relatively versatile program of this sort has been written it remains available as an engineering tool for studying any particular code in detail using long runs and various channel parameters.

The results indicate a consistently large improvement in error rate when even a small amount of interleaving is employed together with retransmission. They also show that interleaving improves the performance of error correcting codes but that the same code provides a better error rate when used with retransmission than when used as a corrector. The increase in redundancy does not appear too

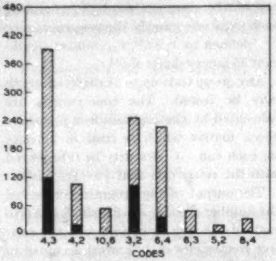


Fig. 1. Observed errors per 50,000 correct transmissions on channel (0.003, 0.034, 0.84)

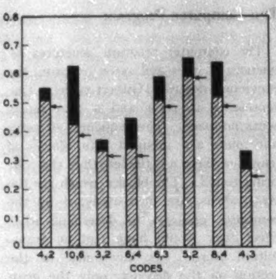


Fig. 2. Redundancy on channel (0.003, 0.034, 0.84)

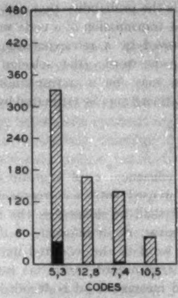


Fig. 3. Relative frequency of errors on channel (0.003, 0.034, 0.84)

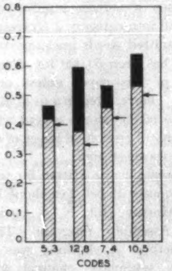


Fig. 4. Redundancy on channel (0.003, 0.034, 0.84)

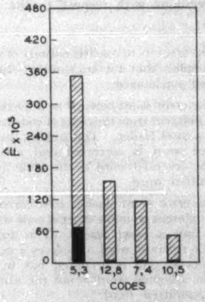


Fig. 5. Relative frequency of errors on channel (0.0015, 0.017, 0.84)

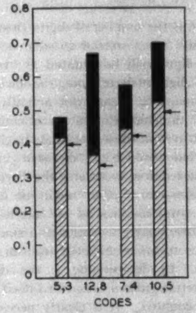


Fig. 6. Redundancy on channel (0.0015, 0.017, 0.84)

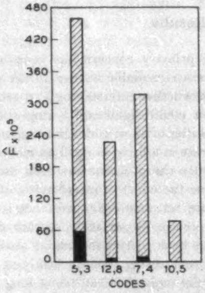


Fig. 7. Relative frequency of errors on channel (0.0144, 0.034, 0.68)

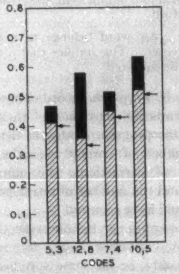


Fig. 8. Redundancy on channel (0.0144, 0.034, 0.68)

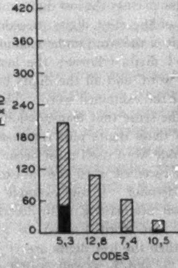


Fig. 9. Relative frequency of errors on channel (0.003, 0.017, 0.913)

severe in comparison with the improvement in error rate but the ultimate engineering judgment on this point may depend on factors that are not considered here.

In making this study, several Gilbert channels were chosen whose error rates varied from bad to worse, which did not represent the error rates on typical commercial channels. Such exceptionally bad channels were selected in order to have many errors for the codes to work on. Even so, the error probability was so small that it cannot be claimed that the relative frequencies of error were good estimates (in a careful statistical sense) of the actual probabilities of error, but only that they were good enough to display the large effect of interleaving and of retransmission over correction.

In all the illustrations the notation

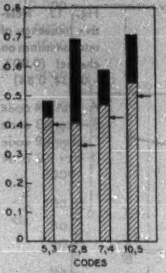
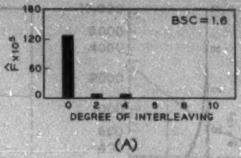
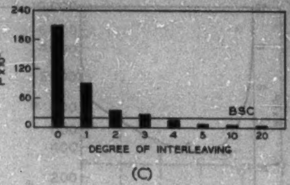


Fig. 10 (above). Redundancy on channel (0.003, 0.017, 0.913)



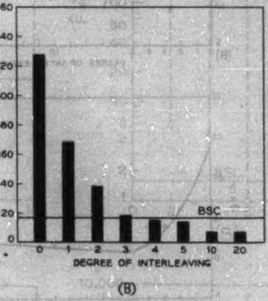
(A)

Using 7,4 code as detector



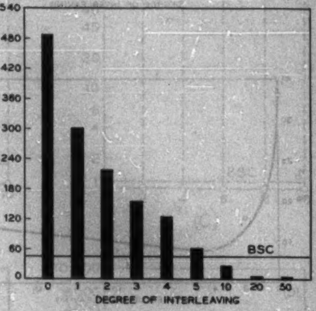
(C)

Using 4,2 code as detector



(B)

Using 4,3 code as detector



(D)

Using 3,2 code as detector

Fig. 11 (right). Relative frequency of errors on channel (0.003, 0.034, 0.84)

(P, π, h) will indicate a Gilbert channel with the given parameters in that order. The notation n,m for codes will indicate the group code that was used of length n with m information places.

In Figs. 1 and 2 a comparison between no interleaving and interleaving at degree 2 for various codes on (0.003, 0.034, 0.84) is displayed. In Fig. 1 the total bar for each code represents the number of observed errors when there was no interleaving and the program was run until 50,000 undisturbed words had been received. The solid part of the bar represents the number of errors under the same conditions except that the code was interleaved at degree 2. In Fig. 2 the effective redundancy for the degree 2 interleaved code is represented by the total bar and the shaded portion represents the effective redundancy when there is no interleaving. The arrow on each bar notes that part of the redundancy due to the check places. For example the 10,6 code has an intrinsic redundancy of 0.4 and the remaining redundancy is due to retransmission.

Figs. 3 and 4 represent a similar comparison on the same channel for the four codes shown except that the relative frequency of errors in the total accepted words is employed in Fig. 3 rather than the actual number of observed errors.

These same four codes are used on three other channels and the results displayed in Figs. 5–10 utilize the same format as in Figs. 3 and 4. The parameters were chosen so that the probability that a randomly selected digit is in error is the same for all channels and is about 0.013. A consistent improvement in error rate was noted when the codes were interleaved at degree 2. The average improvement over the four channels in the error performance of the 5,3 code was more than 5 to 1 and of the 7,4 code about 10 to 1. The interleaved 12,8 and 10,5 codes showed no more than 3 errors in any run.

The redundancy of the codes as shown in Figs. 4, 6, 8, and 10 was greater when the codes were interleaved. The average increase in redundancy was about 13% for the 5,3 code, 68% for the 12,8 code, 20% for the 7,4 code, and 26% for the 10,5 code.

The 10,5 code, without interleaving, and the 5,3 code, interleaved at degree 2, appeared to be comparable both in error rate and redundancy. Longer runs were made with these codes on channel (0.003, 0.034, 0.84). The program ran until 10^6 undisturbed words were recorded for each code, thus simulating roughly 45 minutes to an hour of transmission at 2,000 bits per second for the 5,3 code and about twice that for the 10,5 code. For the 10,5 noninterleaved code 77,298 retransmissions were recorded and 466 words in error gave a relative frequency of error of 46.6×10^{-5} and redundancy of 0.536. The 5,3 interleaved code did slightly better on both counts. The 19,032 retransmissions and 440 words in error that were recorded gave a relative frequency of error of 44.0×10^{-5} and a redundancy of 0.488.

These results point to what are probably the two principal effects of interleaving. The first, which had been men-

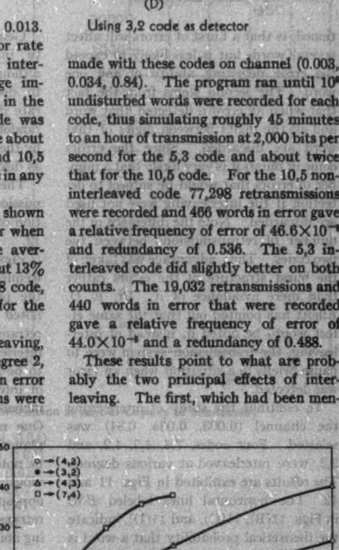


Fig. 12. Redundancy added by retransmission for four codes on channel (0.003, 0.034, 0.84)

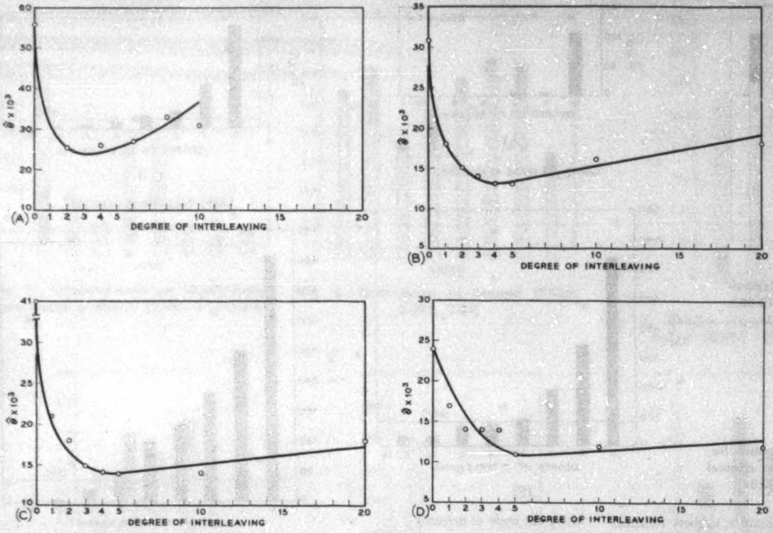


Fig. 13. Relative frequency of retransmissions on channel (0.003, 0.034, 0.84)

A—For 7,4 code
B—For 4,3 code
C—For 4,2 code
D—For 3,2 code

tioned, is that a burst of errors will affect several words but is less likely to exceed the error detecting ability of any one word. The second stems from the fact that when an error is detected, that word and all the digits between digits of that word are discarded. Since the errors come in bursts it is very likely that a number of errors that would be subject to detection if the interleaving were not used are discarded. Note the difference in the number of retransmissions between the two codes. Thus, advantage was taken of the burst structure of the errors rather than mere avoidance of it. Dr. E. O. Elliott has pointed out that the same effect would be noted if, instead of interleaving, the transmission system were to pause briefly in a situation when an error was detected.

To continue the study of interleaving the channel (0.003, 0.034, 0.84) was selected. Four codes, 7,4, 4,3, 4,2, and 3,2, were interleaved at various degrees. The results are exhibited in Figs. 11 and 12. The horizontal lines labeled *BSC* in Figs. 11(B), 11(C), and 11(D), indicate the theoretical probability that a word is in error for that code (noninterleaved) on a binary symmetric channel with the same probability that a randomly selected digit is in error as the channel that was used. The notation $BSC = 1.6$ in Fig. 11 (A) indicates that the same probability for the 7,4 code is 1.6×10^{-6}.

Generally, error performance continues to improve as interleaving increases, although the rate of improvement appears to decrease. The fact that the error rate dips below the theoretical rate for a randomly disturbed channel is explained by the discarding of error digits in a burst as discussed.

The redundancy added by retransmission in these runs is shown as a function of the degree of interleaving in Fig. 12.

Figs. 13(A)–13(D) exhibit the relative frequency of retransmissions observed in the previously described runs. It was noted that there was a sharp drop in retransmissions with a small degree of interleaving, but that there appeared to be a minimum so that higher degrees of interleaving typically caused a slight increase in retransmission probability. One must be cautious about easy explanations of this behavior but it may be noted that the two effects of the previously mentioned interleaving had opposite consequences on the number of retransmissions. When the error detecting ability of the code was enhanced there were more errors detected and consequently more retransmissions. However, when interleaved error digits were discarded the likelihood of retransmission decreased.

Note in Fig. 11(A) that no errors were recorded for the 7,4 code when the degree of interleaving was at least 6. To test this code further, a run was made on channel (0.003, 0.034, 0.84) in which the 7,4 code was interleaved at degree 6 that terminated when 10^6 undisturbed words had been recorded. The 28,005 retransmissions and 7 words in error that were recorded gave an error relative frequency of less than 7×10^{-6}, a retransmission relative frequency of 0.027, and an estimated added redundancy of 0.29.

Turning now to comparisons between retransmision and correction the very simple 4,2 code was chosen first. As an error correcting code, this code will correct 3 of the 4 single digit errors per word. When used on channel (0.003, 0.034, 0.84) at various degrees of interleaving the error relative frequencies are shown by the upper sequence of points in Fig. 14(A). The lower sequence of points represents the same relative frequencies for the same degrees of interleaving using the same code on the same channel but with error detection and retransmission instead of error correction. A line connects each related pair for purposes of association but note that a logarithmic scale is used on the vertical axis. The horizontal lines labeled *BSC* have the same meaning as before, the upper line giving the theoretical value for correction and the lower line that for retransmission.

Interleaving did not seem to have a

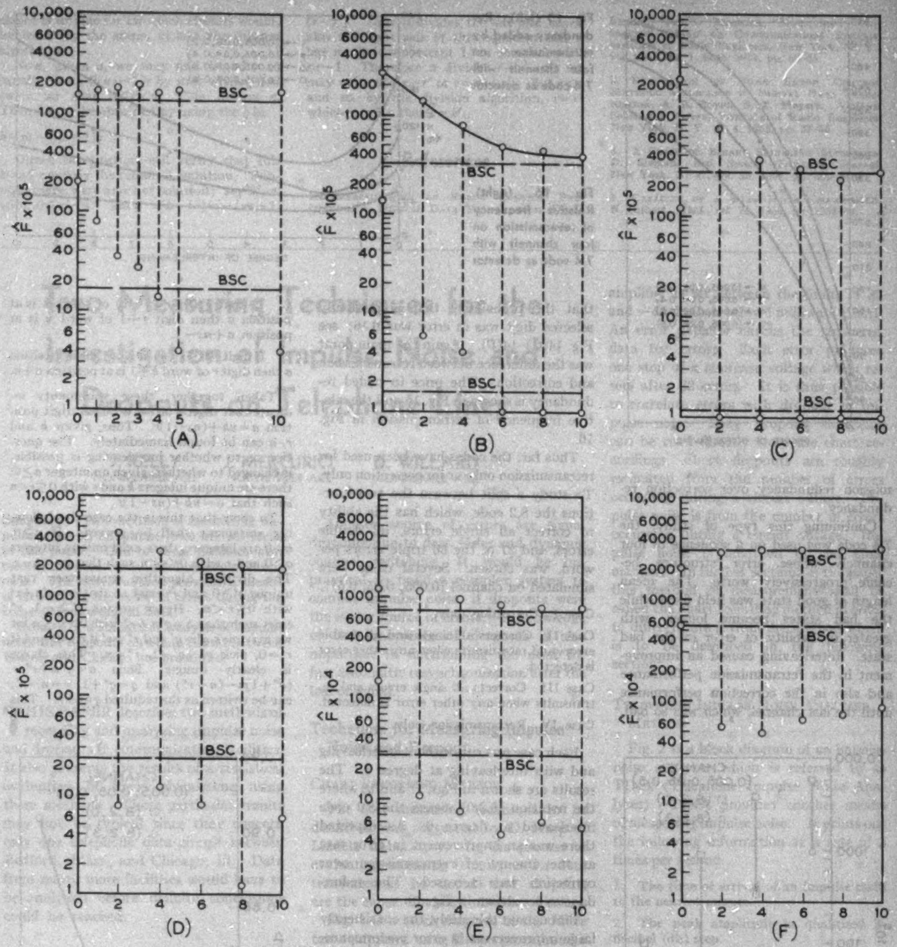


Fig. 14. Comparison of relative frequency of errors for correction and retransmission at various degrees of interleaving

A—4,2 code on channel (0.003, 0.034, 0.84)
B—7,4 code on channel (0.003, 0.034, 0.84)
C—7,4 code on channel (0.005, 0.08, 0.8)
D—7,4 code on channel (0.005, 0.06, 0.6)
E—7,4 code on channel (0.005, 0.04, 0.4)
F—7,4 code on channel (0.005, 0.02, 0.2)

significant effect on this short error correcting code, at least on a channel this bad. It did have an effect on retransmission as has already been noted in Fig. 11(C). But the large difference was between correction and retransmission which differed by one to three orders of magnitude depending on the interleaving.

The 7,4 code was tried under the same circumstances and the results are displayed in Fig. 14(B). Here interleaving had an effect on error correction but again the noteworthy difference was between correction and retransmission.

The price that was paid for this improvement in terms of redundancy is shown in Fig. 12. Added redundancy can be interpreted as the excess of retrans-

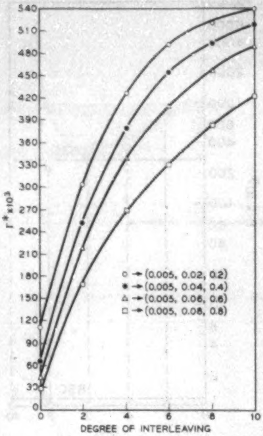


Fig. 15 (left). Redundancy added by retransmission on four channels with 7,4 code as detector

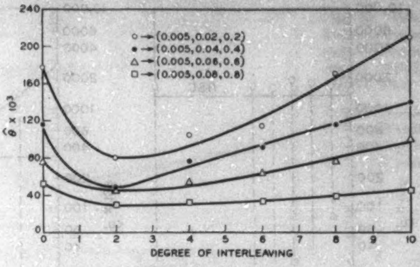


Fig. 16 (right). Relative frequency of retransmission on four channels with 7,4 code as detector

mission redundancy over correction redundancy.

Continuing this type of study, the 7,4 code was used on a sequence of four channels whose error structure became progressively worse. The mean length of good state was held fixed while the bad states became longer, with greater probability of error in the bad state. Interleaving caused an improvement in the retransmission performance and also in the correction performance until the last channel, which was so bad, that the probability that a randomly selected digit was in error was 0.16; see Fig. 14(C)–14(F). Again the main point was the difference between retransmission and correction. The price in added redundancy is shown in Fig. 15 and the relative frequency of retransmission in Fig 16.

Thus far, the codes have been used for retransmission only or for correction only. To study a split between the two functions the 8,2 code, which has the ability to correct all single errors, all double errors, and 27 of the 56 triple errors per word, was chosen. Several cases were simulated on channel (0.005, 0.05, 0.5):

Case I Correction only.

Case II Corrects all single and all double errors and retransmits when any other error is detected.

Case III Corrects all single errors and retransmits when any other error is detected.

Case IV Retransmission only.

Each case was run without interleaving and with interleaving at degree 3. The results are shown in Figs. 17 and 18 where the notation (8,2,0) means the 8,2 code interleaved at degree 0. As expected there was an improvement in error rate as the amount of retransmission over correction was increased. The redundancies are shown in Fig. 18.

Throughout this study, the consistently large improvement in error performance, caused by retransmission combined with interleaving, suggests that this should be pursued as a practical attack on the problem of error control in data transmission.

Appendix

Let the digit positions, in the order in which the digits are transmitted, be numbered beginning with 0. Moreover, let the message words be numbered beginning with 0 and the digits within a word from 0 to $n-1$. Suppose digit 0 of word 0 is at position 0. Then the two statements which define the interleaving are

1. If digit r $(0\leq r<n-1)$ of word k is at position a then digit $r+1$ of word k is at position $a+ns-1$.
2. If digit $r(0\leq r<n)$ of word k is at position a then digit r of word $k+1$ is at position $a+n$.

Taken together, these statements indicate that digit r of word k is at digit position $a=nk+(ns-1)r$. Thus, given k and r, a can be found immediately. The question as to whether interleaving is possible, is changed to whether, given an integer $a\geq 0$, there are unique integers k and r with $0\leq r<n$ such that $a=nk+(ns-1)r$.

To show that this is the case the following statement shall be proved: If $a\geq 0$, $n>0$ are integers, there exist unique integers $q\geq 0$ and r with $0\leq r<n$ such that $a=qn-r$. The division algorithm guarantees that unique $q'\geq 0$ and r' exist so that $a=q'n+r'$ with $0\leq r'<n$. Hence unique $q''\geq -1$, r'' exist such that $a=q''n+r''$ with $0>r''\leq n$ for we may pick $q''=q'$ and $r''=r'$ if $r'\neq 0$ and if $r'=0$, pick $q''=q'-1$, $r''=n$. This choice is clearly unique. Then $a=q''n+r''=(q''+1)n-(n-r'')$ and $q=q''+1$, $r=n-r''$, can be written as the required q and r. This

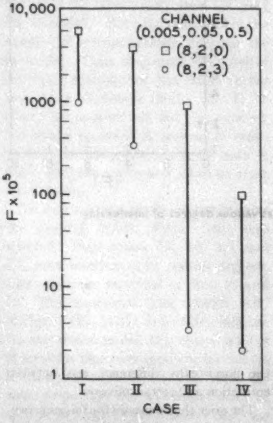


Fig. 17. Relative frequency of errors for various combinations of retransmission and correction

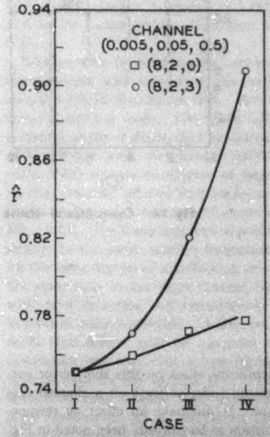


Fig. 18. Redundancy for various combinations of retransmission and correction

choice is unique for two such choices would, by retracing the steps, violate the divison algorithm,.

Now, given a, we may find k and r to satisfy $a=nk+(ns-1)r$ by first writing $a=qn-r$ so that $0\leq r<n$. This defines r. Then k may be obtained by using the q in

$$k=sa-(ns-1)q$$

Direct substitution will verify that this k and r satisfy the required equation. Suppose there were another solution; say k', r' with $0\leq r'<n$. Then $n(k-k')=-(ns-1)(r-r')$. Now n divides the left and hence also the right side of this equation and n has no divisors (except 1) in common with $ns-1$. Therefore n divides $r-r'$ and we may write $nt=r-r'$ or $r=nt+r'$. But $r<n$ and so, by the division algorithm, $r=r'$ which implies that $k=k'$.

References

1. EFFECTIVENESS OF ERROR CONTROL PROCEDURES IN DIGITAL DATA TRANSMISSION, W. R. Bennett, F. H. Froehlich. *Transactions*, Professional Group on Communications Systems, Institute of Radio Engineers, New York, N. Y., vol. CS-9, no. 1, Sept. 1961, pp. 58–65.

2. EVALUATION OF SOME ERROR CONTROL METHODS APPLICABLE TO DIGITAL DATA TRANSMISSION, A. B. Brown, S. T. Meyers. *National Convention Record*, Institute of Radio Engineers, New York, N. Y., pt. 4, 1958, pp. 37–55.

3. A CLASS OF BINARY SIGNALING ALPHABETS. D. Slepian. *Bell System Technical Journal*, New York, N. Y., vol. 35, 1956, pp. 203–34.

4. CAPACITY OF A BURST-NOISE CHANNEL, E. N. Gilbert. *Ibid.*, vol. 39, 1960, pp. 1253–65.

Two Measuring Techniques for the Investigation of Impulse Noise and Dropouts on Telephone Lines

J. KELLY
NONMEMBER AIEE

J. MERCURIO
NONMEMBER AIEE

D. WILLARD
NONMEMBER AIEE

Summary: Two methods of measuring impulse noise on telephone lines are set forth in this paper. One of these methods also records the occurrence and duration of dropouts. The latter method provides a means of identifying the cause of errors during the transmisson of digital data.

Results of a statistical analysis of impulse noise and dropouts on a data circuit between Boston and Chicago are presented.

THIS PAPER describes two methods of recording and analyzing impulse noise and dropouts in communication facilities. It also presents the results of a statistical evaluation of these phenomena using these methods. These particular results may not be typical since they concern only one telephone data circuit between Bedford, Mass., and Chicago, Ill. Data from many more facilities would have to be analyzed before definite conclusions could be reached.

Purpose of Impulse Noise and Dropout Analysis

The causes of impulse noise are well-known, but the amplitudes and distribution of bursts are not yet well characterized. The reason for investigations of impulse noise is obvious since it is a major cause of errors in the transmission of data. With a nearly precise knowledge of impulse noise, a designer should be in a much better position to design data transmission modems that will operate optimumly in the presence of moderate impulse noise.

Another source of errors are signal dropouts due to deep fades and dynamic channel switching. It would be impractical to build a complex system to combat impulse noise, if dropouts were the major source of errors. A knowledge of the occurrence of dropouts is also important in determining the need for fast automatic resynchronization after the loss of signal.

Technique for Measuring Impulse Noise and Dropout

CHART RECORDING METHOD

The first measuring technique to be described was suggested by D. Karp of Lincoln Laboratory. In the process of evaluating smearing techniques for data transmission, he noted that dropouts are the cause of a large number of errors.

Fig. 1 is a block diagram of the Mitre test equipment using Karp's procedure for analysis of dropouts and impulse noise. A 280-cycle tone and the line signal of a Bell data set *201A* are mixed and applied to a telephone line. The composite signal is transmitted from the Mitre building in Bedford to Chicago and looped back to Bedford. At the receive end of the line, the composite signal is passed to the Bell 4-phase data receiver and a tone analyzer. The analyzer consists of an amplifier whose automatic-gain-control (AGC) voltage is a log function of the tone's amplitude. The AGC voltage drives one channel of a 2-channel chart recorder. The dynamic range of the amplifier is set between the limits of 30 and −45 decibels below 1 milliwatt (dbm). An error detector checks the recovered data for errors. Each error produces one step of a staircase voltage which resets after 32 errors. It is thus possible to correlate errors with dropout or impulse noise. Long dropout durations can be read directly from the chart recordings. Short dropouts are roughly estimated from the number of errors occurring during a circuit outage. Impulse noise is from the number of errors occurring during a circuit outage. Impulse noise is analyzed in the same manner. In further sections of the paper this method of measurement shall be called the chart recording method in order to differentiate it from the method which is to be described in the following section.

TELEPHONE IMPULSE NOISE ANALYZER METHOD

Fig. 2 is a block diagram of an impulse noise analyzer which is referred to as TINA (Telephone Impulse Noise Analyser). TINA provides another means of measuring impulse noise. It prints out the following information at a rate of 5 times per second.

1. The time of arrival of an impulse train to the nearest minute.
2. The peak amplitude in quantized 3-decibel (db) step.
3. The envelope duration of a train of impulses.
4. The number of axis crossings included in the envelope.

A dynamic check of the equipment can be achieved using a gated sine wave. Another dynamic check of the amplitude

---

Paper 61-1065, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted June 19, 1961; made available for printing September 6, 1961.

J. KELLY, J. MERCURIO, and D. WILLARD are with The Mitre Corporation, Bedford, Mass.

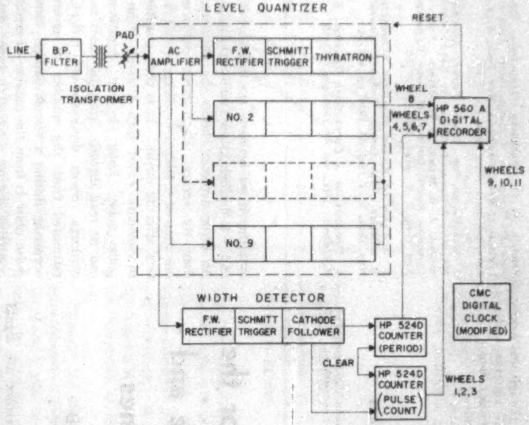

LEVEL QUANTIZER
RESET
LINE
B.P. FILTER
PAD
AC AMPLIFIER
F.W. RECTIFIER
SCHMITT TRIGGER
THYRATRON
ISOLATION TRANSFORMER
WHEEL 8
WHEELS 4,5,6,7
HP 560 A DIGITAL RECORDER
NO. 2
NO. 9
WHEELS 9, 10, 11
WIDTH DETECTOR
F.W. RECTIFIER
SCHMITT TRIGGER
CATHODE FOLLOWER
HP 524D COUNTER (PERIOD)
CLEAR
HP 524D COUNTER (PULSE COUNT)
WHEELS 1,2,3
CMC DIGITAL CLOCK (MODIFIED)

Fig. 1 (left). Block diagram of measuring technique for dropout and impulse noise

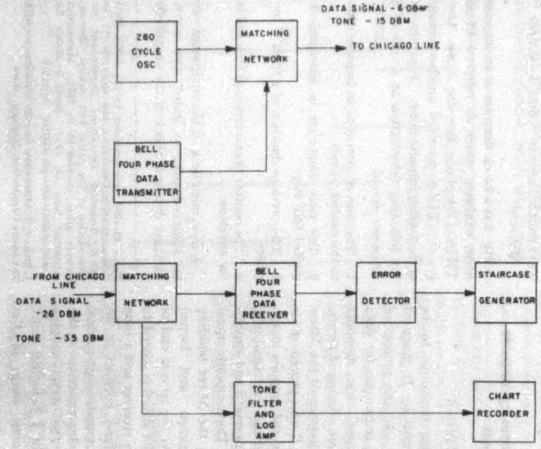

280 CYCLE OSC
MATCHING NETWORK
DATA SIGNAL - 6 DBM
TONE - 15 DBM
TO CHICAGO LINE
BELL FOUR PHASE DATA TRANSMITTER
FROM CHICAGO LINE
DATA SIGNAL -26 DBM
TONE -35 DBM
MATCHING NETWORK
BELL FOUR PHASE DATA RECEIVER
ERROR DETECTOR
STAIRCASE GENERATOR
TONE FILTER AND LOG AMP
CHART RECORDER

Fig. 2 (right). Block diagram of TINA

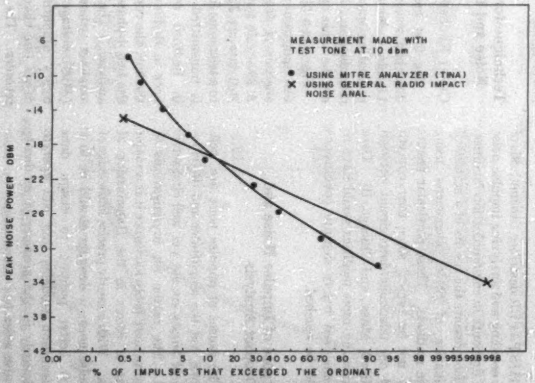


Fig. 3. Impulse-noise analysis of test tape using noise analyzers

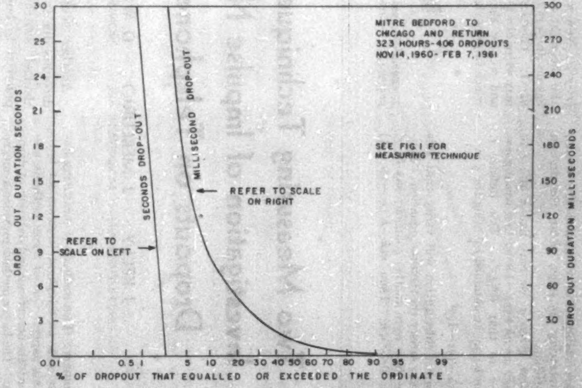


Fig. 4. Distribution analysis of dropouts on Chicago line

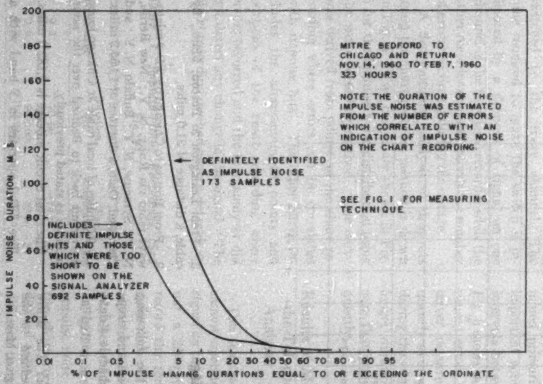


Fig. 5. Impulse-noise duration distribution

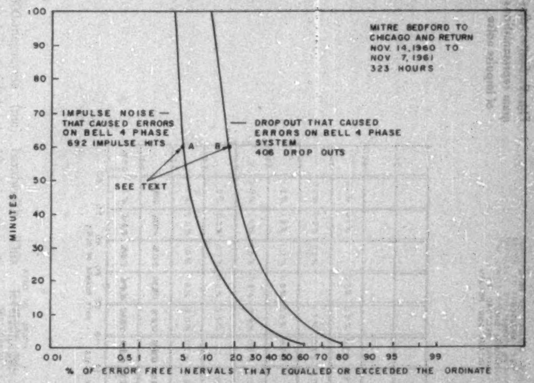


Fig. 6. Distribution of time intervals between errors caused by (1) dropouts, (2) impulse noise

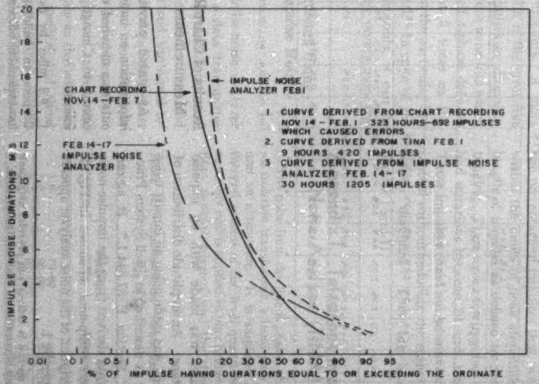


Fig. 7. Comparison of impulse-noise distribution using TINA in one case and chart recordings in the other case

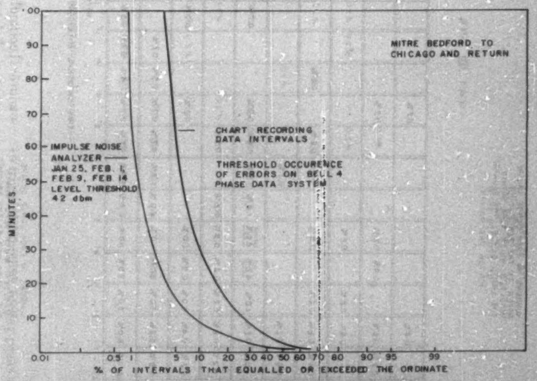
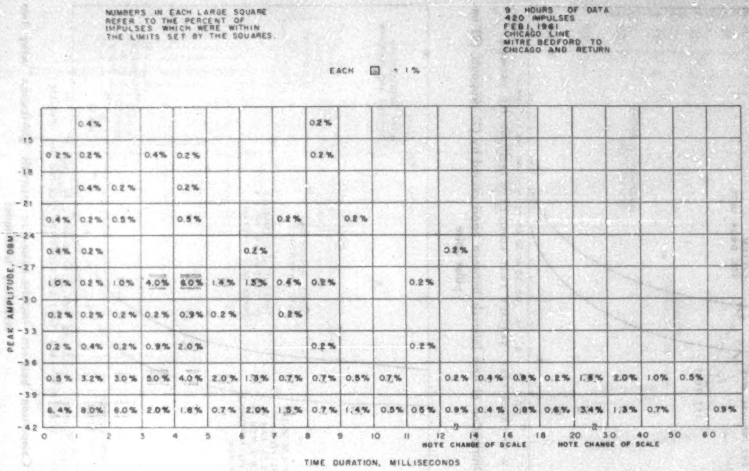


Fig. 8. Comparison between impulse-noise-free intervals distribution using two measuring techniques



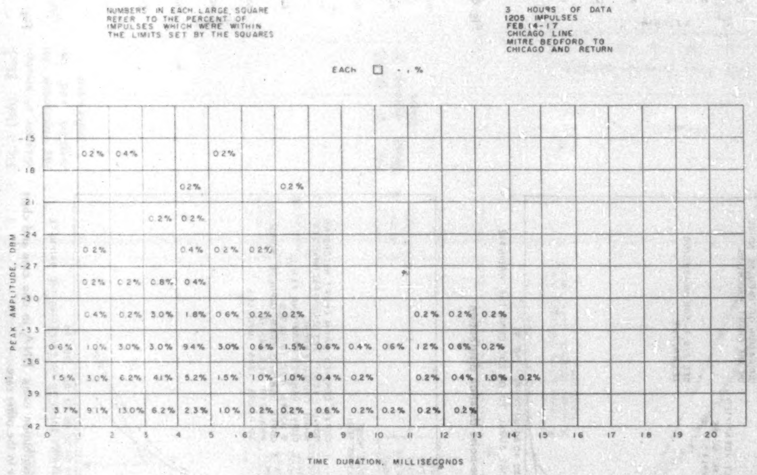


Fig. 9. Scatter diagram representations of impulse noise

quantizer was made with an impulse noise tape. An amplitude distribution of the tape was made using an impact noise meter as described in "Bell System Practices," Section E 37.544.1.* It agreed with data derived from the impulse-noise analyzer equipment as shown in Fig. 3. A description of the noise analyzer appears in Appendix I.

Results of Chart Recording Measurements

As stated previously, the following information on impulse noise and dropouts was obtained on a schedule 4-data circuit to Chicago which is routed as follows:

1. 5.8 miles by cable from The Mitre Corporation building in Bedford to the local telephone office at Lexington, Mass.
2. From Lexington to Boston, Mass., by cable, a distance of 21.0 miles.
3. From Boston to Chicago by *TD-2* microwave radio via Ruby, N. Y.; New Berlin, N. Y.; Syracuse, N. Y.; Buffalo, N. Y.; and Cleveland, Ohio, a distance of 1,055.2 miles.
4. 9.0 miles of local cable at Chicago loops the circuit back to Bedford over the same facilities as stated previously.

The total circuit mileage is 2,182.0 miles.

The *TD-2* operates in *S* band with a

---

* This method is tedious and provides only peak amplitude information on relatively few impulses.

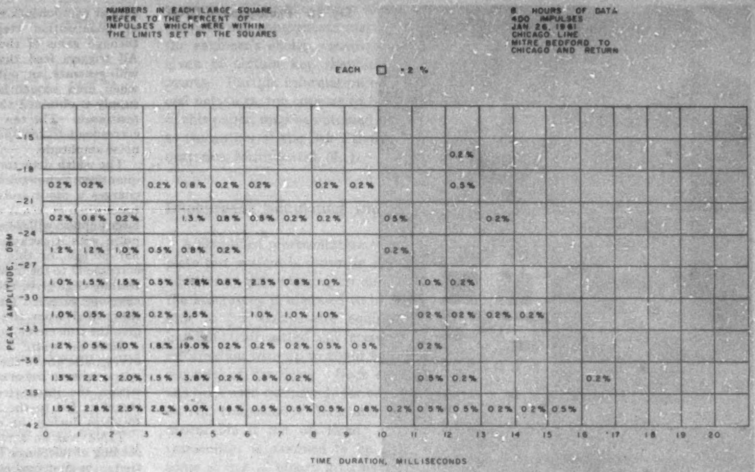


Fig. 9 (continued).

500-megacycle (mc) bandwidth. This band is divided into 12 channels each having a bandwidth of 20 mc and a separation of 20 mc. Each 20-mc channel can carry one-way transmission of 600 voice channels or a one-way television signal. The *TD-2* uses one 20-mc channel as a spare so that if one of the channels experiences a fade, the baseband signal is switched to the spare channel. The time required to make the switch is indicated by a dropout of signal on one channel of the tone level recorder while the other recorder channel records the number of errors occurring during the outage.

A distributional analysis of dropouts using the chart recorder method is shown in Fig. 4. Two curves are shown in Fig. 5 for the impulse noise distribution using the same recording method. One curve was derived from points on the chart recording where errors can be correlated with impulse hits. The other curve includes the aforementioned points and portions of the chart recording where errors occurred and the tone analyzer did not detect any change in the signal. The chart recording method cannot detect low-level impulses of short duration [1–5 microseconds (ms)]. The total dropout time for the 323-hour test period from November 14, 1960, to February 7, 1961, was 4.95 minutes. The circuit reliability for this period is 99.974. The impulse-noise durations totalized over the same interval were 24.7 seconds. This shows that on a long-term basis, the error was determined primarily by dropouts. A more important consideration is the error-free intervals between impulses or dropouts. Fig. 6 shows that impulse-noise hits which caused errors in a Bell 4-phase data receiver occurred about 3 times as frequently as dropouts on the particular line during the test interval. In spite of the longer duration of dropouts, it can be seen that impulse noise occurs more frequently.

Table I. Impulse-Noise Recordings (See Fig. 9)

Percentage	Duration, Ms	Peak Amplitude, Dbm
0.2	70–80	−39 to −36
0.3	80–90	−42 to −39
0.5	90–100	−42 to −39
0.5	100–110	−42 to −39

Results of Measurements Using the Telephone Impulse Noise Analyzer

The telephone impulse-noise analyzer cannot operate with a data transmitter signal on the line. It is possible to transmit a 280-cycle tone on the line and record the tone level and the signal-to-noise ratio of the received tone on a chart recorder. This is done to eliminate impulses which appear during dropouts.

Fig. 7 shows a comparison between the impulse noise durations as derived from the chart recording method and those derived from the noise analyzer during different time intervals. TINA was set to trigger on impulses having peaks above −42 dbm. This is 18 db below the level of the received data signal as used in the chart recording tests. Fig. 8 compares the impulse-noise-free intervals as derived from the two measuring techniques.

TINA can present an interesting picture of impulse noise as shown in Fig. 9. This is a type of scatter diagram showing the percentage of impulses that occurred within the limits of peak amplitude and length of impulse as defined by the rectangles. There were a few impulses which had durations between 70 and 110 ms. These impulses could not be conveniently plotted without reducing the scale of the scatter diagrams. They are listed in Tables I and II.

Appendix I. Telephone Impulse Noise Analyzer (TINA)

The noise from the line is passed through an isolation transformer, a band-pass filter, and an attenuator. It is amplified as an integral process of the level quantizer subassembly and is applied to nine bridge rectifiers, Schmitt triggers, and thyratron cir-

Table II. Impulse-Noise Recordings (See Fig. 11)

Percentage	Duration, Ms	Peak Amplitude, Dbm
1.0	70–80	−42 to −39
1.0	80–90	−42 to −39
0.5	90–100	−39 to −36
0.9	100–110	−42 to −39

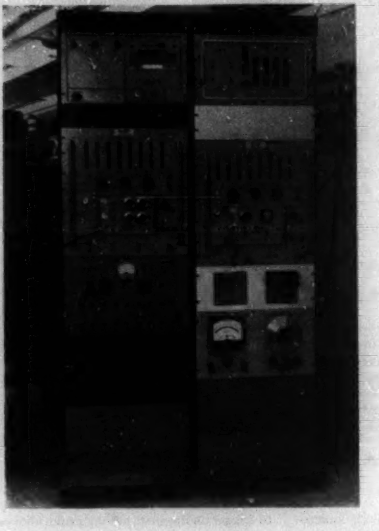


Fig. 10. Photograph of TINA

cuits which generate a staircase voltage corresponding to input amplitude. At the same time the amplified noise is applied to the width detector through a full-wave bridge rectifier. The duration of the detector's output corresponds to that of the pulse chain and is measured in an electronic counter. A second counter receives a pulse from the width detector for every pulse in the input chain and totalizes this information. The output of these two counters, the level quantizer staircase voltage, and the digital clock feed the eleven wheels of a digital recorder. A print command is generated in the width counter at the end of the pulse chain and the printing operation clears both counters and resets the level quantizer.

The subassemblies function in the following manner: The suitably amplified noise is fed to two quantizer stages by a single cathode follower through a split secondary transformer. In each stage the noise is bridge-rectified and superimposed on the bias voltage supplied to the input of each Schmitt trigger. All triggers operate with similar components while the desired range of quantization steps may be obtained through gains of the preceding amplifiers. All triggers feed thyratron circuits which will generate an output staircase voltage when fired sequentially, since their plate supply is obtained through a common load resistance. The ten print-out digits then correspond to ten quantized levels of peak noise amplitude.

The width detector is similar to the level quantizer; however, the trigger now charges a diode isolated resistance-capacitance time constant circuit. A sufficiently large impulse will then generate a flat-topped pulse with a decaying trailing edge. The flat portion of this pulse will, of course, correspond to the impulse duration while the trailing edge gives the machine a chance to envelope detect. The latter time must be subtracted from the readout duration to give the true envelope duration.

Since the width detector trigger fires for every pulse above the lowest threshold, this output gives a count of the number of pulses within the integrated pulse chain and is therefore fed to the second counter which totalizes during the chain duration.

TINA was an attempt at automatic recording of telephone impulse-noise information. As first conceived, it was to record peak amplitude only, however, the additional functions of width, time, and pulse count were added to analyze more adequately the character of telephone impulse noise. TINA should also be a useful tool in analyzing other comparable media such as radio multiplex circuits, voice channels, very-low frequency, etc.

This machine is a laboratory prototype built to test the feasibility of such a system, and specifications as to weight, power consumption, etc., should not be a measure of what could be realized if the required circuitry were designed for its specific function, rather than using existing equipments. (See Fig. 10 for the equipment of the machine.)

Data System Tests Using Simulated Transmission Impairments

F. T. ANDREWS, JR.
NONMEMBER AIEE

Summary: Test instrumentation has been developed for simulating the principal types of transmission impairments that will be encountered on voice-frequency telephone circuits. Such impairments include delay and attenuation distortion, frequency shift, phase and amplitude hits, interfering tones, and noise. This test equipment, together with conventional random-data signal generators and error comparators, makes possible the evaluation of data systems with a wide range of transmission characteristics.

A DATA TRANSMISSION system, complete in two parts, consists of (1) the communications channel and (2) the terminals which prepare and receive the transmitted information. The channel imposes limitations on the system capabilities because of bandwidth, noise, distortion, and other factors. The designer of data systems has little control and often little knowledge of the channel, particularly when its connections are switched through the message network.

Faced with the problem of optimizing the many parameters defining a data terminal, the designer usually turns to extensive field tests over real circuits. Such tests are necessarily time-consuming and poorly controlled because of the variable channel characteristics. The effects of design changes or the advantages of new transmission methods are often difficult to determine.

As our knowledge of the telephone plant characteristics increases,[1] their simulation in the laboratory becomes practical.

Paper 61-1012, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted April 19, 1961; made available for printing August 15, 1961.

F. T. ANDREWS, JR., is with the Bell Telephone Laboratories, Murray Hill, N. J.

Design of the test program and development of suitable simulating equipment have resulted from co-operative efforts of many individuals within the Bell Laboratories. Though impossible to mention them all, special thanks should go to J. R. Davey, Q. D. Groves, A. Lender, R. A. Gibby, and D. W. Nast. B. Jaasma has applied the methods described to obtain the illustrative material.

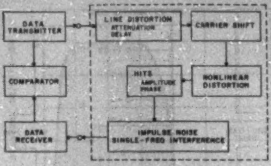


Fig. 1. Standard test equipment

The test system described herein is a first step in this direction, and it should greatly facilitate system designing. Further, by adopting certain standard test conditions and procedures, more meaningful comparisons can be made between the work of individual laboratories.

In delay and attenuation distortion and frequency shift, characteristics can be closely matched over the range reported for real circuits. Since hits, interfering tones, and impulse noise have not been as well characterized, rather special varieties are generated by the test instrumentation, primarily to permit measuring the susceptibility of data systems. The thermal noise that must be added to obtain a given error rate serves as a yardstick in ascertaining the effect of other impairments.

This apparatus cannot in absolute terms predict the performance of systems in the field. However, it can be used to determine system differences and the effects of parameter changes without resorting to time-consuming and poorly controlled field experiments. By standardizing a finite sequence of tests, the relative results obtained in various data-transmission laboratories can be judged more accurately.

Our purpose in describing the test system and procedures is to encourage others to duplicate this approach to the problem. While the design of the more conventional components may be left to the engineer's choice, circuit details are given of certain key simulating equipments. Further information on the special networks, too copious to be included in this paper, may be obtained by writing to the author at the Bell Telephone Laboratories, Murray Hill, N. J.

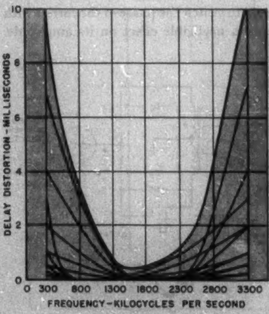


Fig. 2. Delay-simulator characteristics

Impairment Simulating Units

A simplified representation of the complete test system is shown in Fig. 1. In addition to data transmitter and receiver, the laboratory is assumed to have, or be able to get, comparator equipment for handling transmitted and received binary information with up to about 9 ms (milliseconds) of flat delay. Such equipment has been used extensively for error counts on real transmission lines, and is readily procurable if not on hand. The data transmitter is assumed to be driven by some source of binary information, preferably random in character, in order to give the most general transmission results.

Shown within the dashed area of Fig. 1 are the functional units for simulating transmission conditions. Impairments which can be introduced are: so-called line distortion caused by variations in attenuation and envelope delay with frequency; carrier shift commonly produced by single sideband transmission systems; nonlinear amplitude distortion; step changes or hits, both in phase and amplitude; and impulse noise and single-frequency interference. How each of these impairments is produced in the simulating system at the Bell Laboratories is discussed in the next section.

Envelope-Delay Distortion

Both voice-frequency cable and carrier systems introduce significant differences in envelope delay at frequencies within the band used for data transmission.

Knowing the various types of transmission facilities and how they are combined to form built-up connections, an envelope delay-distortion simulator has been built.[2] The simulator networks are designed to produce four families of delay-distortion curves; see Fig. 2. By selecting characteristics from each family, it is possible to closely match any relatively smooth characteristic within the envelope shown. Measurements have revealed that switched-message connections are of this sort.[1]

This is not true for private-line circuits that have been specially conditioned for data transmission. Practical delay-equalization schemes usually produce some residual ripples across the band, which cannot be well simulated with the characteristics of Fig. 2. Furthermore, situations may arise in which a transmission channel is overequalized. To permit simulating such conditions, a group of Western Electric *200*-type telephoto circuit equalizers and a specially designed delay-equalizing network are included in the test system. The *200*-type equalizers (two *AE*'s, two *T*'s, three *AA*'s, one *R*, and one *S*) equalize 5 links of *K* carrier and 150 miles of *H44* cable from 1,000 to 2,600 cycles. The special network more closely compensates this same combination of facilities from 800 to 2,900 cycles, and is a reasonable compromise equalizer for switched-message network connections.

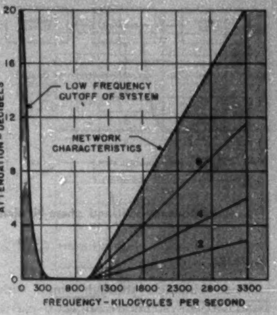


Fig. 3. Attenuation-simulator characteristics

Attenuation Distortion

Based on extensive measurements of the telephone plant, it appears that a suitable approximation to actual transmission conditions can be achieved by a family of linear-attenuation characteristics, together with fixed low- and high-frequency cutoff characteristics.[1] Typically, the low-frequency cutoff affects transmission only when below 300 cycles and the high-frequency cutoff when above 2,600 cycles. Hence, it is the attenuation characteristic between these cutoffs which largely affects practical data-transmission methods.

The test system's family of linear-attenuation characteristics is shown in Fig. 3. Between 1,000 and 2,600 cycles, a difference in loss, ranging from 0 to 14 db (decibels), can be achieved by combining three specially designed networks. While no cutoffs are specifically provided, at least a low-frequency one in the vicinity of 300 cycles is inherent in the transformer coupling of the test system.

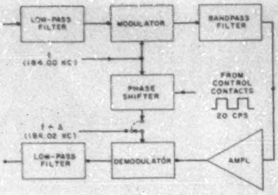


Fig. 4. Frequency-shift and phase-hit simulator

CARRIER SHIFT

Voice channels through single sideband carrier systems are subject to carrier shift, due to frequency differences between modulating and demodulating carriers. While on 99% of switched-message connections, the shift is estimated to be less than 10 cycles, extreme values of 20 to 40 cycles can be encountered. Such a shift, however small, destroys the harmonic relationship between frequency components and, hence, the waveform of a baseband signal. The problem is avoided in practical data systems by transmitting on a carrier within the band. However, carrier shift still affects narrow-band systems and must be taken into account in developing new transmission methods.

Carrier shift is simulated in exactly the same way that it occurs in the field—by modulating and demodulating the data signal with offset carriers as shown in Fig. 4. It was particularly convenient in our laboratory to use Western Electric *O* carrier equipment for this purpose, though other carrier equipment or specially designed circuits can be employed.

Since the *O*-system normally transmits and recovers the channel carrier for demodulation without producing carrier shift, a simple modification is necessary. This involves breaking the carrier recovery circuit and introducing the output of a second independent oscillator. By oppositely misadjusting the two oscillators, shifts up to 80 cycles will result. Unfortunately, the equipment also causes some delay and attenuation distortion, which must be taken into account in planning tests and interpreting the results.

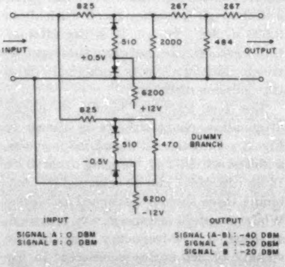


Fig. 5. Nonlinear-distortion simulator

NONLINEAR DISTORTION

Harmonics and cross-modulation products, resulting from nonlinearities in transmission channels, impair the detection of data signals. The Kendall effect, in telephoto nomenclature, means the generation of an unwanted baseband signal which may overlap and be inseparable from the wanted carrier signal.

While Kendall effect can be simulated simply by adding a controlled amount of baseband signal to the carrier signal, a more generally applicable method is to use an accurately controlled nonlinear network. The particular network chosen has a discontinuity in the slope of its transfer characteristic at the origin. The ratio of the slope in the positive and negative directions can be adjusted to give a type of (A-B) cross product, of any desired value, that will produce Kendall effect. Because the discontinuity occurs at zero, the effect of the nonlinearity is relatively insensitive to signal-level variations.

Fig. 5 shows circuitry for producing the nonlinear characteristic, so arranged as to add shunt loss on the negative half-cycles. To maintain a linear input impedance, a dummy section is included to equally shunt the input on positive half-cycles.

HITS

Sudden changes in the amplitude and phase of signals, commonly called "hits," are caused by various disturbances in the transmission facilities. These disturbances may arise from any of a large number of causes—from automatic switching to protection facilities, from routine maintenance operations involving substitution of spare for working equipment, and from momentary failure of a transmission part. While much study is being given to hits, all their causes are not known, and they are not well characterized statistically. In the limit, complete opens in transmission or 180-degree shifts in phase can be experienced.

Though phase and amplitude hits do not usually occur alone, it is most practical to simulate them independently and determine their separate effects. Circuitry for generating hits of both varieties are shown in Figs. 6 and 7..

The amplitude-hit simulator is simply a *T* attenuator with added shunt loss introduced by the contact of a relay operating at 20 cycles. The normal-through-loss is sufficiently high to minimize the effect of the periodic impedance mismatch. The hit rate is vastly more rapid than expected in the field, but serves to give a quick estimate of the amplitude-hit sensitivity of any system under test.

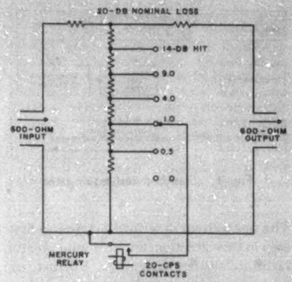


Fig. 6. Amplitude-hit simulator

Phase hits are produced by switching the phase of the demodulating carrier used with *O* carrier equipment, as indicated in Fig. 4. A shift of *X* degrees in the demodulating carrier will shift every component of the signal by *X* degrees. This method corresponds to one mechanism of phase-hit production in real facilities. The phase of the carrier can be shifted abruptly over an 180-degree range by the circuitry shown in Fig. 7. The voltage impressed on the grid of the electron tube is that between the transformer's mid-point and the junction of the resistor and capacitor. Its vector representation is the radius of the familiar circle diagram shown at the bottom of Fig. 7. The angle of the vector can be changed by varying the ratio of resistance to capacitive reactance. If the driving-circuit impedance is low, a single relay contact can be used to switch the phase of the carrier with only a negligible effect on its amplitude.

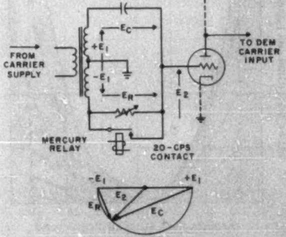


Fig. 7. Phase-hit simulator

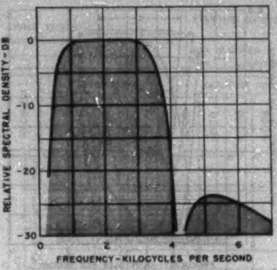


Fig. 8. Random-noise spectrum

A 20-cycle rate again gives a rapid indication of susceptibility to phase hits.

INTERFERING TONES

In a properly operating telephone channel, no tones should be of such magnitude as to disturb a listener—or impair data transmission. Nonetheless, many audible signals, ranging from power harmonics to signaling tones, can appear as a result of trouble conditions. A data test system would not be complete without a means of determining sensitivity to interfering tones. Any oscillator variable from 60 to 3,700 cycles and to a level 20 db above the received data signal level suffices for this purpose.

NOISE

Data systems are subject to noise of a highly complex structure in the field environment. Though progress is being made toward characterizing noise, it is not now possible to establish test conditions for it in the same sense as for delay and attenuation. Instead, noise plays the special role of a yardstick in measuring the effect of other impairments. What is required for this purpose is a source of random and a special form of impulse noise.

The former is produced by a commercial noise generator and frequency-shaped by a low-pass filter, the Western Electric *128A*, which has a 3-kc band as shown in Fig. 8.

For impulse noise, alternate positive and negative impulses are generated by differentiating a 10-cycle square wave using a 250-μsec (microsecond) time constant. These impulses are added to the signal path through the same *128A* low-pass filter. The original impulse and the ringing waveform finally applied are shown in the oscillogram of Fig. 9.

The Test System

The individual functional units of the simulating system have now been described. Their organization into a workable test system is more involved than Fig. 1 suggests. Fig. 10 shows the signal levels, losses, and connection options that have been found suitable.

It is most appropriate to make all the system interconnections at a central jack field at an impedance level of 600 ohms, and to have balanced circuits without grounds at the input and output of each functional unit. Since this leads to using a great many repeat coils in a tandem connection of units, a reasonable compromise is to provide every functional block with at least a floating-transformer coupled output. This gives full flexibility of interconnection, within the limits imposed by levels, and eliminates the possibility of ground loops.

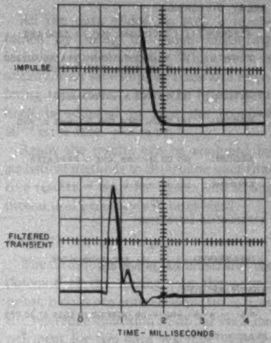


Fig. 9. Impulse-noise simulation

In describing the system organization, the levels are given in terms of the long-time average powers of a continuous random-data signal. As shown in Fig. 10, the output of the data transmitter is brought to about +20 dbm (power in db referred to 1 milliwatt) by means of an adjustable pad and amplifier no. 1. Since the loss and delay simulating units were designed with 1,000-ohm unbalanced im-

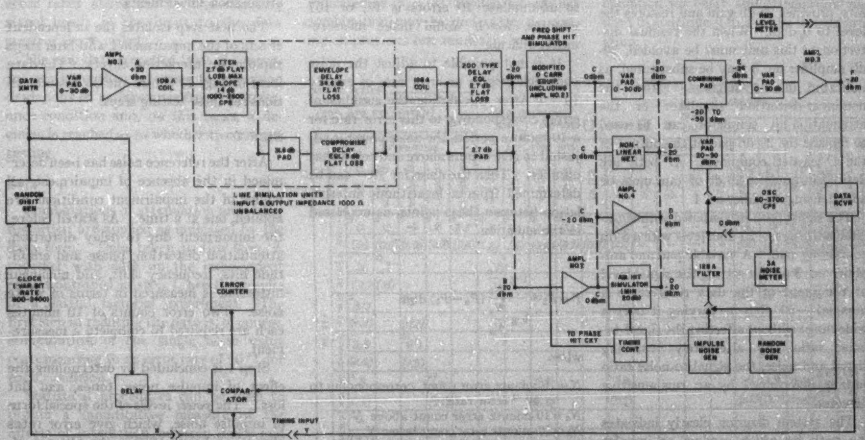

DATA XMTR
VAR PAD 0-30 db
AMPL NO.1
A +20 dbm
108 A COIL
ATTEN 2.7 db FLAT LOSS MAX SLOPE 14 db 1000-2500 CPS
ENVELOPE DELAY 34.6 db FLAT LOSS
31.6 db PAD
COMPROMISE DELAY EQL 3 db FLAT LOSS
LINE SIMULATION UNITS
INPUT & OUTPUT IMPEDANCE 1000 Ω
UNBALANCED
108 A COIL
200 TYPE DELAY EQL 2.7 db FLAT LOSS
2.7 db PAD
B -20 dbm
FREQ SHIFT AND PHASE HIT SIMULATOR
MODIFIED O CARR EQUIP. (INCLUDING AMPL NO.2)
C 0 dbm
VAR PAD 0-30 db
D -20 dbm
C 0 dbm
NON-LINEAR NET.
D -20 dbm
AMPL NO.4
C -20 dbm
D 0 dbm
AMPL NO.2
B -20 dbm
C 0 dbm
AM HIT SIMULATOR (MIN 20 db)
D -20 dbm
TO PHASE HIT CKT
TIMING CONT
COMBINING PAD
-20 TO -50 dbm
VAR PAD 20-50 dbm
E -26 dbm
VAR PAD 0-30 db
AMPL NO.3
F -20 dbm
RMS LEVEL METER
OSC 60-3700 CPS
0 dbm
128 A FILTER
3A NOISE METER
IMPULSE NOISE GEN
RANDOM NOISE GEN
DATA RCVR
RANDOM DIGIT GEN
CLOCK (VAR BIT RATE 600-2400)
ERROR COUNTER
DELAY
COMPARATOR
TIMING INPUT
X
Y

Fig. 10. Data-transmission test system—over-all block diagram

DELAY DISTORTION:

REFERENCE: (1) ± 100 USEC, 300 TO 3200 CPS

SLIGHT: (2) 5 LINKS K, 150 MI H44, EQUALIZED ± 100 USEC, 1000 TO 2600 CPS

MEDIUM: (3) 5 LINKS K UNEQUALIZED

MEDIUM: (4) EQUALIZER FOR 5 LINKS K, AND 150 MI H44

MEDIUM: (5) 20 MI H88, ONE V REPEATER

EXTREME: (6) 10 LINKS K, 300 MI H44

ATTENUATION DISTORTION:

REFERENCE: ±0.5 DB, 300 TO 3200 CPS

IMPAIRMENTS: 2-, 6-, 10-, AND 14-DB SLOPE, 1000 TO 2600 CPS

AMPLITUDE HITS:

0.5-, 1-, 4-, 9-, AND 14-DB INCREASE IN LOSS AT 20 CPS

PHASE HITS:

20°, 50°, 80°, 110°, AND 170° SHIFT IN PHASE AT 20 CPS

FREQUENCY SHIFT:

20-CPS SHIFT IN LINE FREQUENCY

NONLINEAR DISTORTION:

(A-B) CROSS MODULATION 20 DB DOWN ON EQUAL-LEVEL TONES

Fig. 11. Test conditions

pedances, they are connected in tandem between 600- to 1,000-ohm repeat coils. As indicated in the diagram, the delay simulator can be replaced with a compromise equalizer network, together with a flat pad to preserve the flat loss. The standard *200*-type delay equalizers follow the *108A* output transformer and can be replaced by a 2.7-db pad to preserve an output level of −20 dbm.

The modified *O* carrier equipment, used to produce frequency shift and phase hits, provides 20-db gain and restores the level to 0 dbm. When the residual distortion to this unit must be avoided, 20-db amplifier no. 3 may be substituted, as indicated in the diagram. Either the nonlinear-distortion simulator or the attenuation-hit simulator can be used to replace the 20-db pad at the output of the *O* carrier equipment. If both are desired in tandem, 20 db of gain must be added, using amplifier no. 4.

The signal path is combined with simulated noise at a −20-dbm level with a 6-db combining pad. A variable pad and amplifier no. 3 permit setting the signal level at the input of the data receiver to a nominal −20 dbm or varying it over a wide range without affecting the signal-to-noise ratio. By alternately removing signal and noise, the signal-to-noise ratio can be determined by an rms-sensitive detector.

The system diagram clearly indicates the path by which random or impulse noise or interfering tones can be added to the signal. A Western Electric *3A* noise meter or its equivalent can be used to determine an equivalence between random noise measured with message weighting and with the rms-sensitive level meter. Similarly, an oscilloscope can be used to determine an equivalence between the peak amplitude of the special impulse noise and its rms value. Once these relationships have been established, such measurements need not be repeated.

Test Procedures

Using the described system, many tests involving all combinations of impairments can be made. While it is not possible to predict what tests will be needed to solve specific data system problems, it is possible to use a discrete set of tests to bring out the key differences between data systems. Such a program of tests has been established to meet the need for comparisons at Bell Laboratories.

Except for certain special cases, the performance of data terminals is measured in terms of tolerance to thermal noise in a 3-kc band for a given error rate. Hence, thermal noise compares the effects of test conditions and of data systems. As emphasized in the introduction, it will not directly indicate the field performance, because noise of quite a different type will be experienced there.

An error criterion of 1 in 10^4 bits has been selected for these tests. A criterion of 1 in 10^5 or 10^6 would give a more sensitive measure of differences in performance. However, at 1,000 bits per second, the corresponding average time to accumulate 10 errors is 17, or 167 minutes, which would make measurements difficult.

It is not possible to adjust the noise yardstick to give an error rate of exactly 10^{-4}. After calculating the number of errors corresponding to this error rate for a 10-minute period, the noise power is adjusted to give counts above and below that number. Then the noise for 10^{-4} can be determined from a logarithmic interpolation between these points, as expressed in the equation

$$P = P_b + \frac{\log \frac{N_a}{N_b}}{\log \frac{N_a}{N_b}} (P_a - P_b) \text{ dbm}$$

when

N = 10-minute error count, corresponding to 10^{-4} error rate
N_a = 10-minute error count above N
N_b = 10-minute error count below N
P_a = thermal noise producing N_a in dbm
P_b = thermal noise producing N_b in dbm

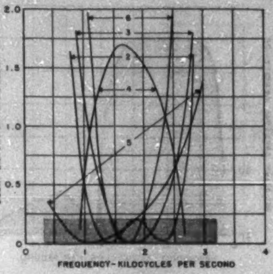


Fig. 12. Delay-distortion test conditions

This interpolation can be performed by drawing a straight line between the two data points plotted with a noise scale linear in db and a logarithmic error scale.

The first step in the testing procedure for a particular pair of data terminals is to establish a reference value of noise. This is accomplished by using the method outlined, with no transmission impairment between the terminals. This is followed by three further steps, in which transmission impairments are applied alone and in combinations. The specific values of selected impairments are listed in Fig. 11 and the delay distortion curves are detailed in Fig. 12. The three steps are to apply:

1. All impairments, one at a time.
2. Selected combinations of delay and attenuation impairment.
3. Selected other impairments, together with a selected combination of delay and attenuation impairments.

The first step isolates the independent effects of the impairments, and later steps explore the interaction effects. The data form of Fig. 13 shows the 35 test conditions for these testing steps.

STEP 1

After the reference noise has been determined in the absence of impairment, all values of the impairment conditions are applied, one at a time. As stated before, the impairment due to delay distortion, attenuation distortion, phase and amplitude hits, frequency shift, and nonlinear distortion is measured in terms of white noise. Two error counts of 10 minutes each are required to complete a measurement.

Step 1 is concluded by determining the effect of impulse noise, tones, and flat loss. The *power* levels of the special form of impulse noise, which give error rates above and below 10^{-4} in the absence of other impairments, are first recorded.

SYSTEM TESTED: ______

STANDARD DATA SYSTEM TEST FORM

DATA RECORDED:
RAW ☐ COMPUTED ☐ DELTAS ☐

DATE: ______
BY: ______
DEPT: ______
LOC: ______

IMPAIRMENT		ATTENUATION					AMPLITUDE HITS					PHASE HITS					FREQUENCY SHIFT	CROSS MODULATION	IMPULSE NOISE	INTERFERING TONE	FLAT LOSS
		NONE (1)	SLIGHT (2)	MEDIUM (3)	LARGE (4)	EXTREME (5)	0.5 DB	1.0 DB	4.0 DB	9.0 DB	14 DB	20°	50°	80°	110°	170°					
DELAY	NONE (1)	0	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	SLIGHT (2)	1	25																		
	MEDIUM (3)	2		26					30					31			32	33	34	35	
	MEDIUM (4)	3		27																	
	MEDIUM (5)	4			28																
	EXTREME (6)	5				29															

Fig. 13. Standard-data system test form

Again in the absence of all other impairments, a single frequency tone is added. The frequency is swept manually back and forth across the band from 60 to 3,700 cycles. The amplitude of the interference is increased until, at some frequency, the error rate is greater than 10^{-4}. The amplitude and frequency of the tone is then recorded as an approximate measure of susceptibility to tones.

Finally, to determine the effect of flat loss, noise is adjusted to the reference level computed to give a 10^{-4} error rate. The level of signal and noise are varied together to obtain a rough measure of the maximum and minimum levels which gives error rates above 10^{-3}. The resulting signal levels at the input to the data receiver are recorded.

The raw data recorded from the measurements are then processed to determine deviations in performance from the reference condition and, on this basis, a decision is reached as to whether to continue testing.

STEP 2

Combinations of attenuation and delay distortion are applied as indicated by the data form. Data taken in this step, 2, together with data on the independent effects obtained in step 1, are the results of a partial factorial experiment. The first step in reducing the data is, of course, computation of the single noise value corresponding to an error rate of 10^{-4}, for each condition.

Statistical procedures can be used to obtain a quantitative measure of the effects of interaction between delay and amplitude distortion. However, an inspection of the results will give some notion of how impairments add, and what practical combinations of distortion can be tolerated. If the data system performs adequately for medium values of delay and attenuation distortion, continuation of the test procedure through step 3 is warranted.

STEP 3

If the third and final step is dictated, a medium-value amplitude and delay distortion is further combined with selected values of phase and amplitude hits as well as with frequency shift and nonlinear distortion. In the data form of Fig. 13, this medium combination is indicated as (3)−(3), though some other combination may be preferable, depending on previous test results.

All the data added by step 3 is obtained by using the same measuring technique as for step 1. White noise is added for all save impulse noise and interfering tone cases. Here the impairment itself is adjusted to give the required 10^{-4} error rate.

Again the results can be analyzed by statistical methods to determine quantitative results. Additive effects and strong interaction probably will be evident.

RESULTS

The data form in Fig. 14 illustrates the outcome of the testing. In the appropriate blocks are entered departures from the reference condition in db for each impairment for two data systems. Entries at the top are for a digital subset, using a 4-phase method of modulation at 2,000 bits per second. Entries in parentheses are for an FM (frequency modulated) subset at 500 bits per second.

For the reference transmission condition, the 4-phase subset required a 10.9-db signal-to-noise ratio for a 10^{-4}-error rate, and the FM subset, a 7.1-db ratio for the same performance. These reference conditions do not appear on the data sheet, but were used for determining the deltas in performance that are shown. The 4-phase subset proves to be more susceptible to delay distortion. In fact, it was not possible to get acceptable error rates, even without thermal noise added, for the (6) delay-distortion condition. Neither system operated for phase hits greater than about 20 degrees. Because of the sensitivity to delay, the (5)−(4) condition for delay and attenuation was used to conclude step 3 of the testing program.

SYSTEM TESTED: 2000 BPS 4-PHASE SYSTEM & (500 BPS FM SYSTEM)

STANDARD DATA SYSTEM TEST FORM

DATA RECORDED:
RAW ☐ COMPUTED ☐ DELTAS ☑

DATE: 9/21/60
BY: E. JAASMA
DEPT: ______
LOC: ______

IMPAIRMENT		ATTENUATION					AMPLITUDE HITS					PHASE HITS					FREQUENCY SHIFT	CROSS MODULATION	IMPULSE NOISE	INTERFERING TONE	FLAT LOSS
		NONE (1)	SLIGHT (2)	MEDIUM (3)	LARGE (4)	EXTREME (5)	0.5 DB	1.0 DB	4.0 DB	9.0 DB	14 DB	20°	50°	80°	110°	170°					
DELAY	NONE (1)	^{0} 0	^{6} 0	^{7} .4	^{8} .8	^{9} 2.7	^{10} .4	^{11} 1.1	^{12} 3.1	^{13} 9.1	^{14} —	^{15} 3.6	^{16} —	^{17} —	^{18} —	^{19} —	^{20} .8	^{21} 0	^{22} 9.9	^{23} 1.4 1.5 KC	^{24} -.51 +[illegible]
		(0)	(0.6)	(1.9)	(3.2)	(5.3)	(.2)	(.6)	(3.2)	(8.7)	(13.7)	(0)	—	—	—	—	(.2)	(.1)	(10.0)	(1.3 1.4 KC)	(-[illegible] +19)
	SLIGHT (2)	^{1} .5	^{25} .7																		
		(.2)	(.8)																		
	MEDIUM (3)	^{2} 9.0		^{26} —					^{30}					^{31}			^{32}	^{33}	^{34}	^{35}	
		(6.2)		(17.7)																	
	MEDIUM (4)	^{3} 3.8		^{27} 6.3																	
		(2.6)		(7.5)																	
	MEDIUM (5)	^{4} .8			^{28} 1.4				4.5			7.3					1.5	1.9	11.4	3.6 1.5 KC	-31 +23
		(.3)			(3.1)				(6.6)			(3.3)					(3.5)	(3.7)	(14.4)	(5.3 1.4 KC)	-41 +19
	EXTREME (6)	^{5} —				^{29} —															
		(9.7)				—															

Fig. 14. Results of tests of two data subsets

Differences in performance, which might be expected between systems with such varied efficiencies in using the transmission medium, are quite clearly brought out by the test program. The results indicate the most productive areas for further detailed study with the view of improved matching of the data terminals to the transmission channels. Only after the design has been optimized by laboratory tests is it necessary to execute tests in order to verify adequacy of performance in the actual field environment.

References

1. Capabilities of the Telephone Network for Data Transmission, A. A. Alexander, R. M. Gryb, D. W. Nast. *Bell System Technical Journal*, New York, N. Y., vol. XXXIX, May 1960, p. 431.

2. A Delay Distortion Simulation Set, Frances C. Dunbar. *AIEE Transactions*, pt. 1 (*Communication and Electronics*), vol. 79, May 1960, pp. 183–85.

A Tone Disabler for Bell System 1A Echo Suppressors

L. F. BUGBEE, JR.
NONMEMBER AIEE

Summary: The capability of simultaneous 2-way transmission over the direct distance dialing (DDD) network will allow data systems to provide more efficient error detection and correction, system control, and circuit assurance schemes. This capability is made available by the development of a tone disabler which will inhibit the suppression function of any echo suppressors in the connection upon command of a conditioning tone and allow normal operation on voice transmission. This paper describes the engineering design and operational features of the tone disabler.

THE DDD network is finding increased use as the transmission medium for data systems. Of particular interest is the teletypewriter exchange service (TWX) scheduled for dial conversion in the fall of 1962.[1] A feature desired by dial TWX and many other data transmission systems is the ability to send signals in both directions simultaneously, an operation precluded by echo suppressors in the telephone toll plant. It is true that transmission can be forced in a direction opposite to that previously established by sending a signal in the reverse direction of sufficient level to "turn-around" the suppressor. In practice, however, the necessary signal levels to reverse the suppressor may exceed plant limits, and the delays involved in turning the suppressor around would certainly complicate signal sequences. By providing the capability for simultaneous 2-way transmission over telephone circuits in the DDD network, more efficient error detection and correction, system control, and circuit assurance schemes can be realized. An echo-suppressor tone disabler has been developed which will inhibit the suppression function of the echo suppressor upon command of a conditioning tone. The engineering design and operational features of this tone disabler are described herein.

The 1A Echo Suppressor

One of the most important factors bearing on the transmission design of the telephone plant is the relationship of return loss performance to the customer tolerance of talker echo as a function of delay time. The accepted engineering practice has been to control short-delay [less than 45-msec (millisecond)] echo by the use of appropriately assigned trunk losses [termed via net loss (VNL) design]. Although this scheme is generally adequate for calls entirely within one regional area, on the longer interregional telephone connections talker echo delays of greater than 45 msec are common, and echo control must be accomplished by use of an echo-suppressor circuit.[2] In contrast to voice design, talker echo is generally of little importance to data systems since 2-way simultaneous transmission is provided by the separation of transmitting and receiving channels. On the other hand, a binary data receiver is usually sensitive to any listener echo delay time greater than a significant fraction of the shortest pulse interval. Echo delay in the telephone plant almost always exceeds this interval for practical data speeds, making VNL design and echo suppressors inappropriate for data systems.

An echo suppressor is a voice-operated device designed for general use in the 4-wire portion of a trunk circuit. A typical 4-wire trunk plan showing the location of the echo suppressor and proposed tone disabler is illustrated in Fig. 1. It may be noted that the disabler is on the drop (office) side of the single-frequency (s-f) signaling equipment and, therefore, free from interference by s-f tones. The two sides of the suppressor are differentially interconnected, making operation dependent on the difference between the speech levels in the even and odd (outgoing and incoming) direction of transmission. For instance, speech power sensed in the even direction of transmission operates a relay chain which introduces a loss of about 40 db (decibels) in the odd direction of transmission. However, a customer may "break in" and cause the suppressor to turn around if he raises his voice enough to deliver a speech power at the suppressor about 5 db higher than the first talker.

The loss inserted in the transmission

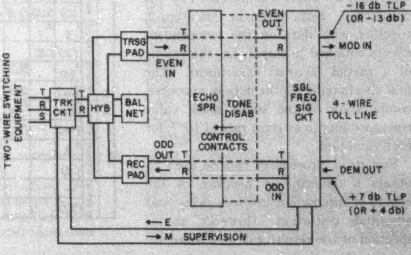


Fig. 1. Typical 4-wire trunk equipment arrangement

Paper 61-1069, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted June 15, 1961; made available for printing August 28, 1961.

L. F. Bugbee, Jr., is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

The author acknowledges the constructive suggestions of D. W. Nast in the preparation of this paper. The disabler circuit design was devised by C. I. L. Cronburg and P. B. Cunningham.

path by the suppressor must remain long enough to prevent the return echo from reaching the talker following a pause in his speech. The length of time that transmission loss remains following cessation of speech power in the opposite path is termed "hangover time" and must exceed the round-trip propagation time of the echo path beyond the suppressor. Tests have shown that best operation is obtained when the hangover time is restricted to less than 70 msec. Where transcontinental or transoceanic circuits are involved, a hangover time exceeding 70 msec would be required to retain effective control of echo if only a single suppressor were employed. To overcome this difficulty, an echo suppressor is located at each end of a long trunk with the suppression function disabled in the incoming direction (odd side) of each suppressor. This arrangement, known as "split operation," permits a reduction in hangover adjustment by an amount equal to the round-trip delay of the trunk. It is necessary to retain differential control so as not to affect the ability of the listener to break in.

The important design parameters for the echo suppressor are outlined in Table I. Note that the suppressor turn-around time is indicated as equivalent to hangover time. Actually, a reversal tone would need to be applied for a somewhat longer time interval in order to account for such factors as propagation time and operating variations.

It is current operating practice to assign an echo suppressor to all inter-regional trunks with VNL larger than 2.5 db, and operate two suppressors on a "split" basis on circuits 2,500 miles and more in length. The number of suppressors installed in the telephone plant is presently estimated at 18,000.

Systems Considerations for an Echo-Suppressor Disabler

The systems design of an echo-suppressor disabler (tone disabler) must take into account the following:

1. Although dial conversion of TWX provides the main impetus for echo-suppressor disabling, the method adopted should allow other data systems which may use the DDD network to disable the suppressors as well. On the other hand, disabling capability should not require undue complications of these other data systems. The time required for disabling the suppressor should not exceed that normally required to complete a switched connection.

2. The disabler must be compatible with existing telephone plant; for example, the probability of disabling tone imitation ("talk-off") by either speech, plant tones, or noise must be minimized while the operating characteristics of the echo suppressor (sensitivity, through and suppression loss, even-to-odd side coupling loss, etc.) must not be compromised.

3. For proper operation, the disabler should be relatively insensitive to expected levels of line noise, whether the disabler is accommodating data or speech transmission on an active trunk or an idle trunk.

4. Certain inherent properties of speech, such as its peaked character and the spectral distribution of energy both on a long-term basis and for syllabic intervals, may be advantageously used to distinguish between a legitimate data signal and speech energy.

With these considerations in mind, a number of schemes could be conceived for the operating method of the disabler. Several of these methods are discussed in the next section, followed by a detailed discussion of the plan finally adopted.

Alternative Methods of Disabling

One plan might be to disable upon receipt of a particular sequential binary code; however, since TWX data sets will far outnumber echo suppressors, it is desirable to avoid complicating these data sets and those of other large customer volume data systems where simplicity and economy are relatively important.

Another plan, termed coincidence disabling, would require the simultaneous receipt of steady data tones from both originating and terminating data sets. The required disabler sensitivity and the return loss performance of the telephone plant, however, make it impossible reliably to differentiate between the reception of two data tones of identical frequency and a single data tone plus its echo. The judicious choice of dissimilar data tones for coincidence disabling would solve this problem and at the same time offer a marked degree of "talk-off" protection. A disadvantage, however, is that during the interval when data tone coincidence is being established, the direction of transmission through the echo suppressor remains ambiguous, thus complicating the control signal sequence required to establish duplex connection.

Disabling upon detection of a signal from either even or odd side of the line (i.e., unilateral disabling), employing a sensitivity and response characteristic commensurate with that of the echo suppressor, and deriving talk-off protection from timing alone is another possibility. An analysis of the statistical distribution of talk spurt lengths of subscribers indicates that to ensure against false operation 99.9% of the time during busy hour would require a protective delay time of from 7 to 10 sec.[3] This delay is unacceptable for many data services in view of the relatively short holding time expected, and, furthermore, the bandwidth of the detection signal would cause vulnerability to noise.

Table I. 1A Echo-Suppressor Design Parameters

Receiver sensitivity (OTLP)	Adjustable between 25 and 35 dbm (db referred to 1 milliwatt) (−31 dbm nominal)
Receiver response	Sensitivity maximum at 1,000 cps (cycles per second), down 20 db at 500 and 2,100 cps
Suppression loss	40 db from 1,000 to 1,500 cps, 30 db from 200 to 3,500 cps
Operate time	12 msec
Hangover time (and turn-around time)	Adjustable between 18 and 275 msec (set to round-trip delay time)

Basic Plan: Tone Disabler

The method of operation selected for the tone disabler incorporates the following features: (1) unilateral disabling control, (2) guard action, (3) talk-off protection interval (attack time), (4) broad-band holding, and (5) timed release.

Disabling of the echo suppressor is accomplished by the continuous and exclusive transmission of energy in a narrow conditioning band (2,000 to 2,250 cps) for a period exceeding the "attack time." As implied, attack time specifies the minimum conditioning tone period required before the disabler transfers to the holding mode and disables the echo suppressor. The principle of guard action, successfully used by s-f signaling equipment, utilizes spectral energy in the voice channel outside of the narrow conditioning band to oppose disabling.[4] Once executed, disabling may be continued by providing energy anywhere in the useful voice band, provided signal absence does not exceed release time. This broad-band holding characteristic will allow broad-voice-band data systems to initiate and maintain disabling, provided only that the proper conditioning signal precedes normal data.

Table II relates the two stable operating states of the disabler to the trunk condition.

Table II. Operating States of Disabler and Relative Trunk Condition

Disabler State	Trunk Condition
Guard mode (normal suppressor operation)	Idle (disconnect) Voice transmission Plant signaling and test tones
Holding mode (suppressor disabled)	Data transmission

Table III. Tone Disabler Design Parameters

Conditioning band sensitivity (OTLP)	Adjustable between 25 and 37 dbm (−31 dbm nominal)
Conditioning band	(2,000 to 2,250) cps
Guard band	(500 to 1,775) and (2,475 to 3,000) cps
Guard-signal ratio	5 db
Attack time	300 msec
Recycle time	10 msec
Release time	100 msec

GENERAL DESIGN CONSIDERATIONS

The successful operation of the tone disabler is contingent upon compromise among several conflicting design factors. To gain a better understanding of the logical selection of design parameters for the tone disabler, several factors are considered for detailed discussion: (1) choice of the conditioning tone frequency and signal bandwidth, (2) guard action, and (3) "attack" and "recycle" times. These factors constitute the principal elements employed to achieve immunity against signal imitation. Other factors such as (4) conditioning band sensitivity, (5) holding mode release characteristics, (6) release time, and (7) noise susceptibility warrant equal attention. Disabler design parameters are summarized in Table III and discussed in detail in the following.

1. *Conditioning Tone: Signal Bandwidth*

The dial TWX data set transmits data on a frequency-shift basis, the mark and space frequencies having a 200-cps separation. It is necessary that the standard voice band accommodate at least two narrow-band channels to allow for duplex operation. The center frequencies ($F_1 =$ 1,170 cps and $F_2 = 2{,}125$ cps) of the narrow-band TWX channels have been carefully selected (*a*) to avoid the greater transmission losses and distortion encountered at the upper edge of the voice band, (*b*) to be compatible with the direction of propagation of existing plant supervisory and signaling tones, and (*c*) to avoid the s-f signaling frequencies of 1,600, 2,400, and 2,600 cps. The originating (calling) station will always send on the F_1 channel, while the terminating (called) station will send on the F_2 channel. The TWX connect sequence requires that the mark frequency of the upper channel (F_{2M}) generated at the called station shall be the first data signal transmitted during the course of call establishment and the presence of F_{2M} at the echo suppressor will cause disabling of that device.

The conditioning tone is thus set at $F_{2M} = 2{,}225$ cps. The probability of talk-off is very strongly influenced by the choice of conditioning tone frequency, decreasing rapidly as this frequency is increased. Data obtained from tests made during the development of in-band s-f signaling circuits are reproduced in Fig. 2. It is apparent that 2,225 cps is a fortunate selection of conditioning tone for talk-off minimization.

Freedom from talk-off is also enhanced by using the least possible signal bandwidth, thus effectively reducing the power of speech and noise falling within the conditioning band. Adequate bandwidth must be provided, however, to encompass the signal frequencies of existing and proposed data systems which will benefit from the suppressor disabling feature. Additional latitude must be allowed for frequency shift encountered in nonsynchronous single-side-band carrier systems, and variation of generated signal frequencies and component values in the tuned circuits of the disabler.

A conditioning band that brackets the range (2,000 to 2,250 cps) at nominal sensitivity has been selected.

2. *Guard Action*

The most significant deterrent to talk-off is the use of practically all voice-band frequencies outside of the narrow conditioning band to generate a voltage which is used to oppose that derived from the conditioning band.

The vowel sounds contribute the major portion of total speech power. The phonetic specification of a vowel involves the fundamental frequency (100 to 250 cps) and the "formants" which define the regions of energy concentration produced by vocal resonances.[5] These vowel properties are subject to wide variations caused by the word or sentence context in which the sound is spoken, and the age, sex, or dialectal background of the speaker. Fig. 3 illustrates the average position of the formants and their relative magnitudes for vowel sounds contributing significant second formant energy within the conditioning band. Formant energy tends to diminish with increasing frequency; therefore, response shaping at

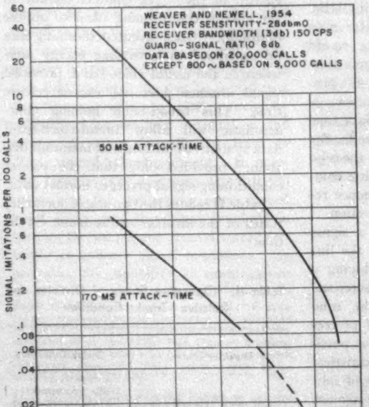
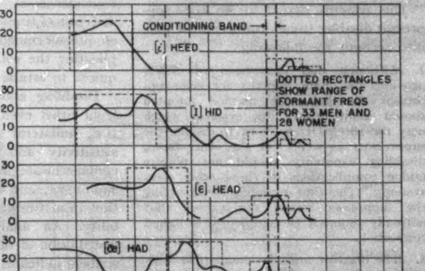


Fig. 2 (left). Conditioning-tone-imitation data

Fig. 3 (below). Typical vowel spectra contributing significant energy in the conditioning band (2,000 to 2,250 cps)



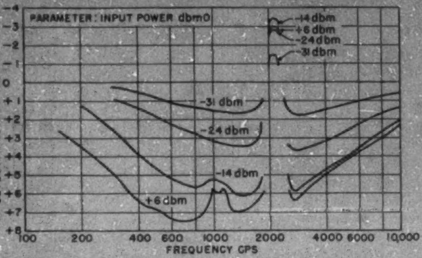


Fig. 4. Detector characteristic

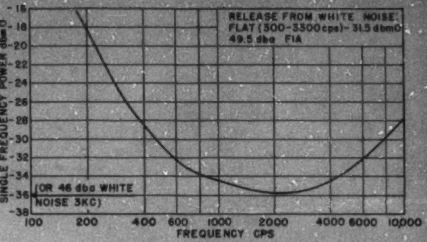


Fig. 5. Release characteristic (tone applied to only one direction of transmission)

the low end of the guard band must be tailored to utilize fully first formant energy to prevent false disabling by these sustained vowel sounds. Guard action above 2,500 cps may take advantage of the upper formants and also protect against very-high-level speech by sensing modulation products caused by amplifier overloading.[6,7] A plot of the resultant direct voltage as measured at the output of the disabler detector is shown in Fig. 4 as a function of frequency; the several contours represent constant s-f input powers referred to the 0 transmission level point.

3. *Attack Time and Recycle Time*

Talk-off protection can be improved by providing the longest possible attack time consistent with data-set timing and switching sequences required to establish the connection. Basically, this suggests a time greater than the average syllabic period (125 msec) and noise burst lengths. An attack time of 300 to 400 msec meets this requirement, while providing a margin for disabling, even under noisy conditions, within the 690-msec minimum timing interval required by the TWX connect sequence. Talk-off advantage is enhanced and the reliability of tone disabling improved by making the attack-time interval independent of the input signal magnitude.

To satisfy the disabling criterion, a comparison must be made between the actual time interval during which only conditioning tone has appeared and the required attack time as derived above. Any deviation from the "tone-only" condition should immediately reset the timing mechanism. Resolution is a function of the recycle time, which should be brief enough to prevent integration over a time period spanning normal consonant or vowel intervals, otherwise a train of tone bursts or speech syllables could cause false disabling even though no single burst could do so independently.

4. *Conditioning Band Sensitivity*

The transmitting levels of data sets are set as high as possible to secure an adequate signal-to-noise margin at the receiver; however, crosstalk, intermodulation distortion, and repeater overloading impose practical limitations. The dial TWX output level is adjusted to compensate for most transmitting loop loss variations, and provides a −10-dbm signal (at 2,125 cps) at the switch of the serving central office, although for the longest loops this level is reduced to as low as −15 dbm.

The design sensitivity of the disabler in the conditioning band must account for the probability of multilink connections, the expected variation in data-set output levels, tolerance in the performance of the disabler, the higher transmission loss at 2,125 cps, the design and operating variations in trunk loss, and flexibility to adapt to future changes in the plant. A sensitivity of −31 dbm referred to the 0 transmission level point (or −31 dbmO) within the conditioning band should permit a dial TWX data set at the end of the longest loop to initiate disabling for 99% of the switched connections reaching an echo suppressor on final routes in the DDD network.

5. *Holding Mode: Release Characteristic*

Once disabling has been established by the recognition of a legitimate conditioning tone, it must be sustained by energy anywhere in the useful voice band. Holding sensitivity must, therefore, be designed to accommodate expected data signal levels at the echo suppressor. In addition, it must be made at least equal to that of the conditioning band to prevent hunting (i.e., mode instability).

The disabler must release from the "holding mode" upon sensing an idle trunk condition, in the face of the maximum line noise that may be expected consistent with commercial-grade telephone service. This places an upper limit on holding sensitivity, and a compromise must be sought between the two opposing considerations mentioned above. Fortunately, such a compromise has been found; a plot of the minimum s-f tone level needed to sustain holding is plotted versus frequency in Fig. 5. Special attention must be given to the holding response where a concentration of noise power can be expected from harmonics of 60-cps or stray plant supervisory tones.

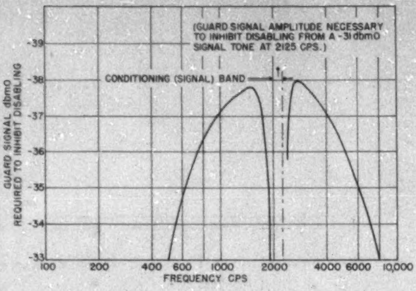
6. *Release Time*

The release time should span any momentary transmission dropout (signal loss in both even and odd side of the 4-wire line is necessary for release) and provide sufficient latitude so that future data systems can easily adapt their signal sequences to sustain disabling. The release must be fast enough to cause a minimum of impairment upon accidental speech disabling or false holding on a noisy trunk. A timed release of 100 msec is the design objective for the tone disabler.

7. *Noise Considerations*

Having derived the minimum required sensitivity for the conditioning band and related the sensitivity of the guard band in a manner that will minimize talk-off, the probability that noise power falling in the guard band might delay or prevent conditioning tone recognition must be explored. With a signal-guard characteristic as shown in Fig. 6, total random noise power (flat noise in a 3-kc band) of −34 dbmO will extend attack time by 5%, while −31.5 dbmO of noise will inhibit entirely disabling of a 2,125-cps conditioning tone at −29 dbmO. As mentioned previously, the disabler must also release from the holding mode in the presence of the maximum expected trunk noise. The holding release characteristic of Fig. 5 will permit the disabler to transfer to the guard mode with a total noise power input of −31.5 dbmO.

Available telephone plant noise sta-



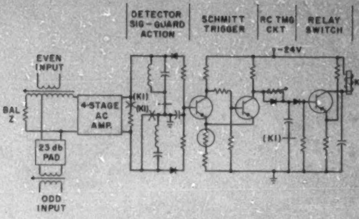


Fig. 6 (left). Guard-signal characteristic

Fig. 7 (above). Tone disabler simplified schematic

tistics are generally given in terms of dba or dbRN (units which employ frequency weighting designed to measure relative interfering effect to a listening subscriber) and do not indicate the spectral distribution of noise energy or even the rms noise power. Assuming flat noise within a 3-kc voice band, a line noise of −32 dbmO or greater can be expected at the disabler only 1% of the time.[8] However, since both even and odd transmission paths are sampled, the disabler will incur some noise penalty.

Circuit Design Description

The tone disabler includes a hybrid mixing circuit, a 4-stage transistorized a-c amplifier, a tuned differential detector, a Schmitt trigger, an *R-C* timing circuit, and a transistor switch controlling a relay which disables the suppressor. A simplified diagram is shown in Fig. 7.

The hybrid combiner couples the signals from both even and odd sides of the transmission path to the a-c amplifier, while maintaining an isolation loss between even and odd sides of at least 50 db in the voice band. The line-bridging impedance is very high so as not to interfere with echo-suppressor operation. Two pad values are available to compensate for the different transmission level points that may exist at the even and odd paths.

The a-c transistor amplifier consists of two 2-stage feedback amplifiers in tandem, and has a thermistor-compensated feedback loop to improve temperature stability. A gain-control potentiometer allows sensitivity adjustment of about ±6 db. A ripple filter is used to render the amplifier insensitive to battery noise.

The detector circuit employs an anti-resonant tuned network to develop conditioning signal voltage and a series tuned network to generate guard voltage. These voltages are half-wave rectified, filtered, and polarized so as to oppose each other. A direct relationship between disabler operation and multiple-input signals is shown in Fig. 6 where the minimum s-f guard signal necessary to inhibit disabling is plotted against frequency for a constant conditioning tone (2,125 cps) applied at −31 dbmO. Relay contacts control the broad-band holding feature by effectively replacing both tuned circuits with a resistive load. Choice of this load resistance determines the holding mode release characteristic which allows disabling to cease when s-f energy in the voice band drops below −35 dbmO.

The Schmitt trigger has a negative threshold which is temperature-compensated by a thermistor. An input signal in the conditioning band (2,000 to 2,250 cps) of −31 dbmO or greater will cause the Schmitt to operate. On the other hand, with a −31-dbmO tone at the center of the conditioning band, and a −36-dbmO tone applied in the guard band from 700 to 1,775 cps or from 2,475 to 3,000 cps, the Schmitt will not trigger.

The operation of the Schmitt trigger initiates a charging cycle in the *R-C* timing circuit. When the voltage developed across the capacitor exceeds the breakdown voltage of the avalanche diode plus the reverse bias on the transistor, the switch will conduct and close the relay. The attack time can be varied by adjusting the timing resistance in series with the capacitor. If the trigger should revert back to normal before the operation of the relay, the capacitor is quickly discharged through the diode shunting the adjustable timing resistor, thus securing a minimum recycle time.

The release time of 100 msec is achieved by the following: A back contact of relay *K1* removes the capacitor from the timing circuit, thus preventing loss of charge upon a momentary signal dropout that would necessitate recycling through the attack-time interval to reestablish disabling, and the collector of the relay switch transistor is bypassed to ground by a large capacitor to delay the relay release as well as minimize the inductive kick of the relay.

The disabler has been designed to operate from the standard 24-volt positive ground signal battery, used also by the echo suppressor. The maximum current drain is approximately 30 milliamperes. The disabler will operate successfully over an ambient temperature range of 30 to 140 degrees Fahrenheit.

The existing *1A* Echo Suppressor Test Set can be used to line up and maintain the tone disabler. In addition, attack and release times can be measured.

Initial tests of the disabler, covering the range of signal levels that can be reasonably expected at the suppressor, produced no signal imitations from normal speech or program material. A more extensive testing program is in progress which will give a broad statistical sampling of disabler performance as it relates to both voice and data transmission.

References

1. A View of Things to Come, **W. A. Stevens.** *Telephony*, Chicago, Ill., vol. 159, Oct. 22, 1960, p. 84.

2. Nation-wide Switching of Intertoll Trunks—Transmission Considerations, **O. F. Wallman.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, May 1958, pp. 223–30.

3. Characteristic Time Intervals in Telephone Conversation, **A. C. Norwine, O. J. Murphy.** *Bell System Technical Journal*, New York, N. Y., vol. 17, Apr. 1938, pp. 281–92.

4. Single-Frequency Signaling System for Supervision and Dialing over Long-Distance Telephone, **N. A. Newell, A. Weaver.** *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 489–95.

5. Toward the Specification of Speech, **J. C. Steinberg, R. K. Potter.** *Journal of the Acoustical Society of America*, New York, N. Y., vol. 22, no. 6, Nov. 1950, pp. 807–20.

6. Technical Aspects of Sound (book), **E. G. Richardson.** Elsevier Press, Inc., Houston, Tex., 1957, pp. 199–231.

7. Control Methods Used in a Study of the Vowels, **G. E. Peterson, H. L. Barney.** *Journal of the Acoustical Society of America*, vol. 24, no. 2, Mar. 1952, pp. 175–84.

8. Capabilities of the Telephone Network for Data Transmission, **A. A. Alexander, R. M. Gryb, D. W. Nast.** *Bell System Technical Journal*, vol. 39, May 1960, pp. 431–76.

The 81-A Exchange Trunk Carrier System

MYRON E. FERGUSON
ASSOCIATE MEMBER AIEE

MAURICE C. HARP
NONMEMBER AIEE

THE post war expansion pattern of the telephone industry has created a substantial demand for carrier equipment suitable for exchange trunk connections. There is an increasing demand for these office-to-office connections in extended area service (EAS). It has been the cost rather than technical feasibility that has discouraged extensive use of carrier equipment in the exchange plant. However, if a carrier system is available at a price competitive with cable cost, the demand will support the sort of manufacturing volume that is associated with mass production techniques. It has been the primary design target of the 81-A carrier system to prove-in economically at distances as short as 2 to 3 miles, and thus a substantial market potential is insured.

Each basic 24-channel system requires two ordinary voice-frequency (v-f) cable pairs, one for each direction of transmission. Four such 24-channel assemblies can be mounted in a single $11^1/_2$-foot by 19-inch equipment rack. In Fig. 1, a 24-channel assembly of this equipment is illustrated, which shows how the channels are separated into two 12-channel subgroups with common equipment mounted between. The carrier and power supply at the top of the illustration will handle up to four such 24-channel terminal shelf assemblies. The necessary power supply to power the cable repeaters is included in the common equipment so that no intermediate repeater power feed is required, even for systems 15 miles in length.

A performance advantage that is obtained by the use of carrier over cable pairs is indicated in Fig. 2. The use of medium to long exchange trunks, or the tandem connection of trunks, tends to run into losses great enough to affect the grade of service being supplied to customers. Customarily, v-f repeater amplifiers are added to cable pair trunks so that connection losses will be held to a minimum, but the use of the carrier eliminates this problem. The 2- to 4-db (decibel) via-net-loss (VNL) rating of the carrier channelizing permits operating all trunk circuits at a standard electrical length. In the end, customer connections are substantially independent of distance, and are closely grouped about an optimum VNL.

Another matter of concern in the introduction of carrier into the exchange plant is that of maintenance and reliability. There is no question that the v-f cable pair requires little maintenance, except the occasional routine check of the trunking relays. The carrier system which is to replace cable pairs must also require little maintenance, or the initial economic advantage gained will be spent in maintenance during the life of the equipment. Fortunately, components of proved reliability are now available which will yield the desired continuity of service. However, in the design of this system, a rather substantial quality assurance program must be rigorously enforced. Semiconductors were once an item of some concern. On the basis of accelerated life testing it is difficult to extrapolate far into the future, but careful evaluation indicates that both diodes and transistors, if conservatively chosen, show excellent prospects of a negligible failure rate for many years of normal operation.

Regulation

In order to correct for cable variations with temperature, automatic correction of slope and loss is required. A pilot tone is transmitted just above the highest frequency channel. At the receiving end of the circuit, slope regulation information is derived by comparing this pilot with the sum power of the channel carriers. Variations in slope are only a few db, and are corrected in a single terminal slope regulator. Flat loss variations are corrected at the terminal and at regulating repeater points. Repeater regulation is controlled from the receiving terminal by small variations in the line feed current

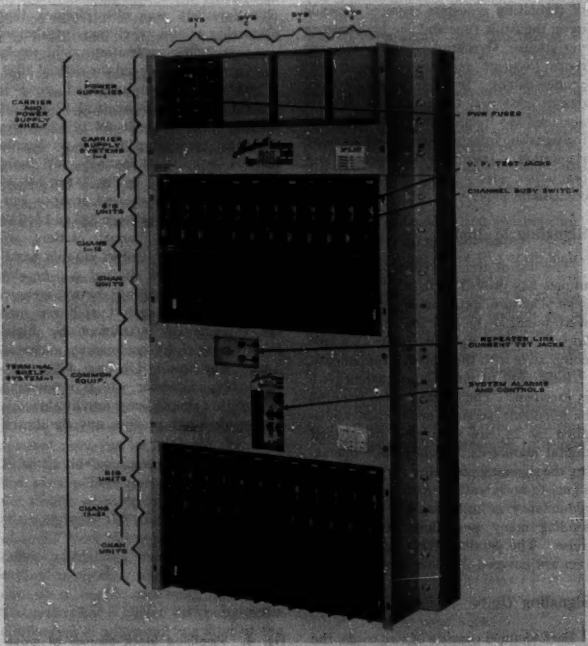


Fig. 1. The 24-channel terminal with carrier generator and power regulator

---

Paper 61-774, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 20, 1961; made available for printing May 25, 1961.

MYRON E. FERGUSON and MAURICE C. HARP are both with Lenkurt Electric Company, Inc., San Carlos, Calif.

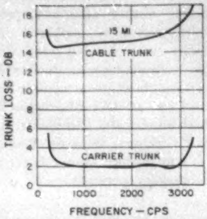


Fig. 2. Via net loss of carrier channel and 15-mile physical trunk

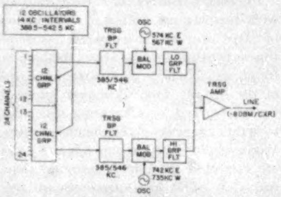


Fig. 3. Simplified block diagram of group modulation

which powers the repeaters. The number of regulating repeaters required depends upon the cable type and the temperature variations. Overhead cable requires more frequent regulation than underground cables. In the interest of system design economy, the regulating pilot is also used as a frequency-modulated carrier for signaling information as well as slope regulation. Details of the repeaters and regulation are covered in a separate companion paper.[1]

Signaling Multiplex

The use of time division digital techniques has proved increasingly useful in several areas of data transmission. They were found to be practical in providing 24 independent signaling paths at a cost compatible with the system requirements.

Reliability of multiplex signaling path is excellent. The pulsing rates required in telephone service, even when multiplied by 24, are very slow by present digital standards, which permit operating the signaling path at a very favorable signal-to-noise ratio. Also, a very high redundancy is built into the system by sending many repetitions of each dial pulse. The details of this signaling system are covered in a companion paper.[3]

Signaling Units

Each channel consists of two cards, the modem unit and the signaling unit. There are four different basic signaling units available to meet the various requirements. For step-by-step exchange trunks, a one-way loop-dial unit is used. This option is comparable to the usual one-way arrangement, which transmits dial pulses by breaking the loop current and provides answering supervision by battery reversal. The E and M signaling unit is compatible with various toll practices, or can be used with existing two-way trunking facilities. A revertive signaling unit is available that will permit trunk signaling between panel offices and certain crossbar offices. A foreign-exchange option takes care of connection options outside the usual office-to-office practice. Where in-band tone signaling is needed, a simple bypass unit is supplied.

Modulation Plan

Each individual channel employs conventional double-sideband transmitted-carrier modulation. In a system design that is strongly based on economy with reliability, it was apparent early in the planning that the more sophisticated approaches were incompatible with cost objectives. It was also obvious that channel compandors, channel regulation, or other such complexities would surely increase the cost beyond the desired level.

The 24-channel transmitted group is generated from two identical 12-channel banks. Thus, a single set of 12 carrier generating oscillators is able to supply carriers for up to 96 channels, or 8 channels per oscillator. In Fig. 3, the way in which the basic 12-channel group is generated in the band between 385 and 546 kc is shown. The channel carriers are spaced 14 kc on centers in this band, with each channel allotted a 7-kc slot leaving 7-kc guardbands between channels. To insure against spillover into the interchannel guardband by high-frequency speech components, a simple 3.5-kc single-mesh low-pass filter precedes the modulator. It is interesting to note that simple linear amplitude modulation does not produce any significant spurious modulation products except at multiples of the carrier frequency. Since the basic 12-channel modulation group occupies less than one octave of bandwidth, no individual channel transmit filters are required, but a single 12-channel bandpass filter 385 to 546 kc provides all the necessary filtering for a 12-channel group.

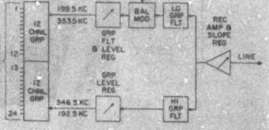


Fig. 4. Simplified block diagram of group demodulation

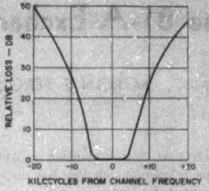


Fig. 5. Typical channel filter response

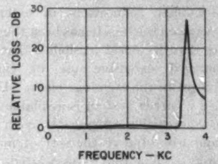


Fig. 6. Receiving v-f low-pass filter response

Each of the 12-channel groups is modulated down in frequency by a group modulator, so that a continuous spectrum of channels at the final line frequency spectrum of 20 to 355 kc is produced. The 14-kc spacings of the channel carrier frequencies are at alternate odd multiples of 3.5 kc, with the East-West channels interlaced between the West-East channels. This particular arrangement reduces certain distortion and crosstalk problems in an optimum manner. For example, the East-West and West-East circuits do not pose a crosstalk problem because any coupling introduces interference only into the interchannel guardband. Most of the intermodulation and harmonic distortion products in the common equipment do not directly coincide with any other channel frequency. Initial planning was undertaken utilizing odd harmonics of 4 kc since this would have been most compatible with past carrier practice. However, this set the maximum channel frequency so high that it would have been necessary to limit the system to 20 rather than 24 channels if cable crosstalk objectives were to be met. Since frequency compatibility between this and other carrier systems was not a basic requirement, the 3.5-kc spacing was adopted. Because signaling tones are

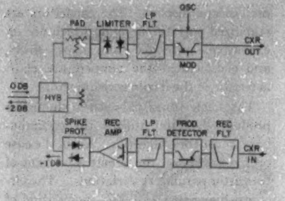


Fig. 7 (left). Block diagram of channel unit

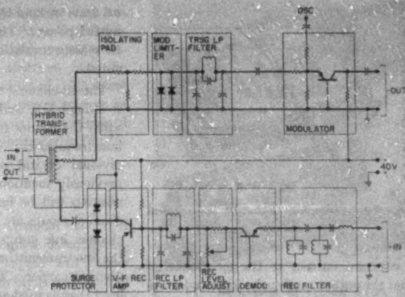


Fig. 8 (right). Simplified schematic diagram of channel modem

not associated with the individual channels, a pass-band extending up to 3.2 kc is achieved within the ±3.5 kc allocated to each channel.

The use of 12-channel group modulators provides a desirable desynchronization of the various channels because they reduce intelligible crosstalk. The channel demodulator uses a carefully linearized product detector that reduces the interfering effect of cochannel crosstalk signals which are not in absolute synchronization with the normal carrier. Although a cursory analysis would indicate that the crosstalk demodulated in this manner will be essentially unintelligible, careful listener tests show that the interfering effect of expected cochannel crosstalk levels is improved only about 8 db.

The demodulation pattern employed at the receiving terminal is shown in Fig. 4. The low-frequency group of 12 channels is modulated up in frequency to approximately the same band as the upper 12 channel group, but it is interlaced so that potential crosstalk will be avoided. The receiving filters in the channels units that separate adjacent channels are made up of three-mesh networks. The guard-bands between the channels, in addition to the back-up of a simple one-mesh low-pass v-f filter, permit such filter simplicity. The linear demodulation of the product detector also contributes to reducing filter requirements, since out-of-band signals are not demodulated into the voice band that is passed by the v-f low-pass filter. The response of a typical channel filter and receiving low-pass filter are illustrated in Figs. 5 and 6.

Circuit Arrangements

The input to a channel in a normal trunk circuit is taken from the second selector in a step-by-step dial office. At this point, the circuit consists of three leads, the tip, the ring, and the control or sleeve lead. The voice and dial loop both appear on the tip and ring pair, and the idle or busy logic information, that is required by automatic switching gear, appears on the control lead. These three leads enter a channel signaling unit where the dial loop is separated from the voice path. At this point the dial loop and control circuits may be dismissed from our consideration and the v-f path through the equipment should be followed. The voice pair feeds a coil-type hybrid which serves to split the transmitting and receiving directions. The hybrid is an asymmetrical unit that affords much greater loss to the transmitted than to the received signal.

The block diagram in Fig. 7 is useful for an understanding of the circuit sequence into the modem unit, and Fig. 8 is a schematic diagram of the basic modem circuitry. Following the hybrid, the transmitted voice path travels through a 5-db pad to insure that the 2-wire line impedance does not follow the impedance variations of the limiter. The limiter consists simply of a pair of unbiased silicon junctions across the signal path. The signal level and impedance level at the limiter diodes have been proportioned to set the cut-in level of the limiting action.

The necessity for limiters in the circuit arises because, with amplitude modulation, there is a definite modulation linearity break point. This is the limit commonly referred to in AM (amplitude-modulated) practice as 100% modulation. Overmodulation would result in so-called sideband splatter, which cannot be tolerated without employing a transmitting filter in every channel to keep the splatter noise out of adjacent channels. In single sideband practice, such limiting has always been present, though it is not usually obvious due to the overload characteristics of the balanced modulators. Here too, splatter is a problem unless a sideband transmitting filter is utilized to obtain a single sideband signal out of a double sideband modulator. The arrangement chosen here permits the elimination of a transmitting filter at the cost of a pair of diodes.

The difference between the limiting in this equipment and that customarily found in toll carrier practice is that it takes place at a lower level. In this system, the common equipment load is almost entirely that of the carriers, with negligible increase in total power, even when all channels are simultaneously active. To keep the carrier load within reason, the modulation level has been set as high as possible without affecting speech characteristics due to limiting. The limiters are just beginning to clip in the presence of a milliwatt (mw) test signal at the input. For this reason, it is desirable to make lineup adjustments lower than the usual test levels so that the potential error associated with the close limiting characteristics could be avoided. A lineup test set is available that will supply 1,000 cycles at −10 dbm (db referred to 1 mw) and will read channel lineup levels in terms of net loss with this signal.

In Fig. 9 the modulation limiting effect of the limiter circuit is illustrated. It can be seen that approximately 1/2 db limiting occurs with a mw signal at the 2-wire input. However, less than 1% of telephone users talk loudly enough to be appreciably affected by this limiting.

The typical channel distortion is shown in Fig. 10. The third harmonic is principally due to the speech limiter.

The modulator circuit uses a grounded-base connection of a single transistor. The carrier supply is arranged to deliver

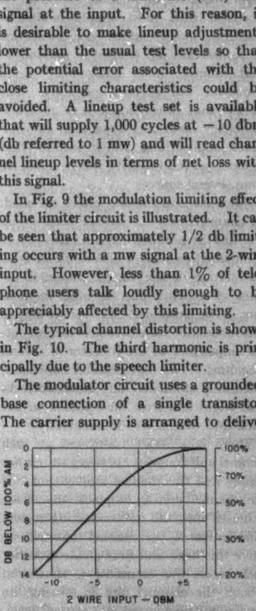


Fig. 9. Characteristic of limiter

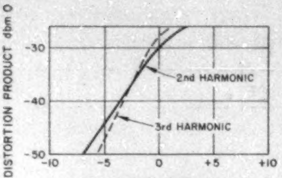


Fig. 10. Harmonic distortion of typical channel

a constant current to the modulator so that the output tends to be independent of transistor parameter variations. The transistor is not biased to pass any d-c current, so the collector output current consists of half-wave pulses of the carrier. When a v-f signal is added to the carrier, the carrier shifts the conduction point of the input junction in accordance with the v-f signal. Thus, the conduction angle of the carrier is varied to yield a net resulting amplitude modulation of the fundamental frequency component of the carrier in the collector current. This modulator circuit is very simple, and is suitably linear for a modulation factor greater than 90%. This represents a higher degree of modulation than is practical with most simple modulator circuits. The grounded-base connection is so tolerant of transistor parameters that almost any small-signal p-n-p transistor will perform perfectly. The device that was chosen is a medium-frequency switching type used in computer applications, which insures that units of proved integrity will be available from several sources.

Since no transmitting filters are used, all channels of a 12-channel basic group are tied in parallel. With many types of modulators, this would result in cross modulation of the various channels because of feedback into the other modulators. The grounded-base connection of this modulator transistor provides an output impedance at the collector of several megohms. This high collector impedance essentially eliminates any signal currents that might flow from the common output back into the various modulators.

There is a diode connected at each channel unit connector between the carrier supply and the channel unit output. If the channel unit is in place, this diode is biased into the nonconducting region. Elimination of the channel unit will remove this bias, and thus create a carrier path across the connector so that even though the channel unit is missing, its carrier is transmitted at normal level. This feature permits the removal of channel units without changing the total transmitted power. This total power is used for system regulation purposes, as will be noted.

The 12-channel basic group is then fed to a single-sideband group modulator that shifts the carrier frequencies to the final line frequencies. This unit has an individual crystal oscillator so that the previously mentioned carrier desynchronization will be insured. Two such 12-channel groups are added to form the 24-channel line group and are amplified in the line transmitting unit to a level of −8 dbm per carrier. This is a slightly higher level than is used at line repeaters, and is provided to insure a safe level margin above office switching noise which might otherwise be troublesome. This level is sufficiently high to permit the use of a first repeater spacing of one and one-half times the normal spacing. This is an economy, since the end repeater section would otherwise be only a half-section long, corresponding to the half-loading section of the v-f loading pattern.

At the far end of the circuit, which is the receiving end of the group of channels under discussion, the last repeater section is usually a half section so that the received signal level into the central office is held high enough to avoid the office switching noise problems. A special high-level repeater is available and is normally used to maintain the high signal-to-noise ratio into the office. The repeaters and the associated level regulation and cable equalization have some very interesting and important features which contribute much to the practicality of the system.

The received group of channels is amplified in a common amplifier which includes automatic regulation of the flat level and automatic equalization of the line-slope variation due to temperature effects. These regulation features are related to the repeater regulation and equalization, and are covered in a separate paper.[1]

The received channels are split into two 12-channel groups with filters. The higher frequency 12-channel group is fed directly to the demodulators, while the lower frequency group is modulated up to the same frequency range as the upper frequency group before it is demodulated.

The receiving channel filters are a part of the channel units. As previously indicated, a simple three-mesh net is sufficient when considered with the guard bands and v-f low-pass filter. The filter uses inexpensive ferrite pot cores which have trimming slugs to permit tuning the final filter on a sweep generator with oscilloscope display. The capacitors are stabilized polystyrene units which are inexpensive, quite free from aging effects, and complement the temperature characteristic of the ferrite core inductors.

The single transistor used in the linear product detector utilizes the common-base transistor connection. As in the case of the modulator, the circuit is tolerant of transistor parameter variations. The circuit is linear to above 90% modulation of the received signal even at relatively low carrier input levels. This is not true of many types of detectors, particularly in the region of modulation troughs where negative-peak clipping is a problem. Low-level diode detectors are basically square-law devices and would be a poor choice for this application. When the signal that crosstalks into the interchannel guardband is considered as an example, it is noted that the channel filter provides only slight attenuation to this disturbance. A square-law detector would demodulate the v-f information from such an interfering carrier at a level corresponding to the level of the carrier of the crosstalk. Measurements on the detector used in this equipment show that such interference is suppressed by 6 db more than twice the crosstalk coupling loss.

Even-order intermodulation distortion and harmonic products generated in the common equipment result in spurious products at 3.5 or 7-kc from a desired carrier. The v-f low-pass filter interposed between the detector and the v-f amplifier has an infinite rejection point centered at 3.5 kc to prevent any 3,500-cycle whistle from ever being audible in the channel output. This filter also rejects the 7-kc whistle that might otherwise be audible from crosstalk into the the interchannel guardband.

The v-f amplifier uses a single transistor in a grounded-emitter configuration. This is very straightforward, and uses local feedback to achieve the required stabilization. This stage is less tolerant to transistor parameter variations than the modulator and demodulator, but even so, the device current gain factor (beta) can age to less than half of normal before circuit degradation results.

A noise generator is provided as part of the system, and is connected into the receiving amplifier so that masking noise may be added if desired. There are certain application situations where such noise is desirable in order to render any crosstalk unintelligible. The use of the noise is optional at the time of system lineup, and the exact considerations involved are covered in a companion paper by L. R. Cool and R. G. Walker.[3]

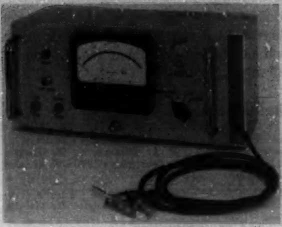


Fig. 11. Portable battery-operated v-f test set

System Alarms

The high-frequency pilot which is used in conjunction with the slope regulator and signaling is also used as a sensing pilot which will cause a major alarm if failure occurs; the disappearance of all channel carriers will also cause a major alarm. Failures in the signaling equipment which would affect all channels are also detected and connected back as a major alarm condition. Failure of 12 channels only brings in a group alarm, but leaves the other 12 channels in normal operation. In case of alarm there are circuit features that will disconnect all the affected customer connections and make all dead trunks busy at both ends of the system. Each individual signaling unit is equipped with a test (butt-in) jack and busy-out key and the signaling unit connector includes a feature which will busy the connected circuit if the signaling unit is removed from its socket.

Test Unit

The only adjustment associated with the channel units is the setting of the receiving level to provide the desired net loss. This is normally set to 2 db for toll connecting trunks, although there is an adjustment range of from 0 to 6 db. A correction of 0 db is not considered practical when used with switching equipment, though there could be special situations where the two-wire line impedance would be precise enough to permit the 0-db adjustment.

In order to set and check the drop level of the channels routinely, a special test set is available. This is connected to the channel via the test jack. When the test set is connected, a lamp indicates whether the channel is busy or idle. When the channel is idle, throwing a switch will busy the channel and feed a 1,000-cycle reference test signal into the channel. An identical unit at the far end is switched to record the received drop level on a self-contained meter that has an expanded db scale. After the level is set in one direction, each operator reverses his test set switches and the opposite direction is adjusted. This test unit is independent of power sources as it is operated by self-contained standard flashlight cells. It is illustrated in Fig. 11.

Mechanics

Each 24-channel terminal of 81-A carrier is a single complete assembly except for the power supply and carrier oscillators. The assembly is composed of six shelves which are handled as a unit. The card-type plug-in units utilize stitched wiring, which is well suited to the 81-A carrier since the majority of the plug-in units can take advantage of the economies that can be achieved with this technique in high volume production. A simple finger hole in the card itself eliminates the need for a bail for removing the cards from the shelf and no front plate is necessary on any of the cards except the alarm unit. A single front cover for the 24-channel assembly is used for dust protection and for a more finished appearance. The cover has cut-outs for in-service access to the busy-out keys and test jacks on the signaling cards and for access to the alarm unit which contains lamps and alarm cut-off switches.

A channel unit and a signaling unit are shown in Fig. 12. It can be seen that the staples stitched into the card serve as tie points for components and for wire jumpers. The leads of the components are inserted through holes in the card adjacent to the staples and the leads are wire wrapped to the staples. Insulated wire leads are run across the card on the component side and bare wire straps are placed on the other side. The stitched wiring technique includes the use of automated equipment for driving the staples into the cards for component insertion, for connection of the bare wire straps, and features automatic wrapping of wires and component leads. A dip solder step completes the automatic wiring of the card. The last manufacturing step includes the insertion and soldering of any components which might be damaged by the heat of the dip soldering operation. Components and staples are individually marked on the cards to permit easy identification and full cross reference to the equipment drawings and maintenance instructions. Maintenance soldering is handled in the usual manner with no need for concern about possible damage to the card or the wiring.

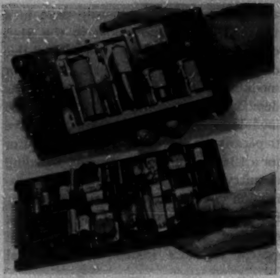


Fig. 12. Mechanics of typical 81-A card unit showing stitched wiring

Where heavy components or relays must be supported by the cards, units are built using two cards with the heavy parts between them. This two card technique is also useful for getting more than the standard 15-socket pins on a unit. No plug hardware is required on the card, since the plug pins are formed of a gold-plated spring stock and inserted automatically by the stitching machinery. The cards are edge mounted in the shelf assemblies and are supported on nylon-slide runners to avoid any binding in the slides. These mechanics result in a rather high density assembly which is made practical by the small amount of heat to be dissipated per unit. The complete elimination of vacuum tubes is very helpful in this respect.

There is considerable heat to be dissipated in certain components associated with the power regulation units. For this reason, the repeater power unit is located at the top of the rack, and the high heat items (the 40-volt regulators) are heat sunk directly upon a large aluminum back cover plate behind the channel banks. This plate spreads the heat out over several square feet of area having unobstructed contact with the room air for free and efficient heat exchange.

Conclusions

The initial requirement on the system for low cost without sacrifice of reliability and performance has been achieved by a concentrated effort by applications, development, and production engineers to insure that every possible cost saving has been included in the final product. Perhaps the most important concept is that the channel equipment, of which there are 24 per terminal, must be as simple as possible even though this means added complexity in the common equipment that appears only once in a terminal. The innovations in the 81-A carrier which were developed around this concept include the single transistor modulator, the single transistor linear demodulator, keeping all of the initial modulation frequencies within an octave to permit the use of a single transmit filter for a 12-channel group, common signaling circuitry rather than separate transmission circuits for each channel, simplified repeaters particularly the regulating repeaters, and a common sophisticated system regulator instead of individual channel regulators. Not only has this approach reduced the cost of the system, but also results in a smaller number of active circuit elements per channel which will assure a minimum of maintenance.

Applications engineering concern dictated the inclusion of the trunk signaling relays as part of the carrier equipment rather than requiring separate trunk signaling equipment. The resulting integration of trunk signaling relays with the carrier system effects a sizeable reduction in overall installed cost per trunk channel. And finally, the development and reduction of the stitched wiring concept for use in the 81-A carrier yields an economic manufacturing technique for equipment of this type.

References

1. The 81-A Exchange Trunk Carrier System—Repeaters, E. A. Gilmore, C. G. Griffith. *AIEE Transactions* (see pages 606–10 of this issue).

2. The 81-A Exchange Trunk Carrier System—Signaling, E. F. Tuck, Donald K. Melvin. *Ibid.* (See pages 610–17 of this issue.)

3. The 81-A Exchange Trunk Carrier System—Applications, L. R. Cool, R. G. Walker. *Ibid.* (See pages 617–23 of this issue.)

The 81-A Exchange Trunk Carrier System—Repeaters and Regulation

E. A. GILMORE
NONMEMBER AIEE

C. G. GRIFFITH
ASSOCIATE MEMBER AIEE

A BLOCK DIAGRAM of the 81-A 24-channel exchange carrier system shown in Fig. 1 indicates the basic system parameters. It is designed for a maximum length of 15 miles and powered by the 48-volt office battery. The frequency allocation is shown in Fig. 2.

Other aspects of the 81-A system are covered by three companion papers in this issue by M. Ferguson and M. Harp, D. Melvin and E. Tuck, and L. R. Cool and R. G. Walker.

Repeaters

The repeaters must compensate for cable losses as illustrated in Fig. 3. Repeaters are installed approximately every 6,000 feet for 22-gage cable at the load points. The frequency response of 13 miles of 22-gage cable including 11 repeaters is illustrated in Fig. 4. While the repeater equalizer is designed to be used with 22-gage cable, it can be used with 19, 24, or 26 by spacing the repeaters at the 400-kilocycle 20-decibel (db) loss points. Since the system is 4-wire, the repeaters are unidirectional.

In Fig. 5 the block diagram of the nonregulating repeater is shown. The input and output transformers isolate the equalizer and gain stages from the line and provide a simplex connection for the power to operate the amplifier. Power for the repeater is derived by the voltage drop across a Zener diode connected between the center points of the input and output transformers. This voltage is filtered by a capacitor before use. In the nonregulating repeater the zener voltage (with ±10% tolerance) is nominally 22 volts; in the regulating repeater it is 25 volts; and in the high level repeater it is 36 volts.

The input transformers match the line to the equalizer. The equalizer furnishes the appropriate impedance at low frequencies where the cable attenuation is a minimum and reduces reflection ripples in the response. At high frequencies the equalizer impedance increases as a result of the parallel mesh. Combinations of *L*, *C*, and *R* (inductance, capacitance, and resistance) are provided to be adjusted for proper high-frequency response. Three continuous adjustments are provided. The potentiometer designated LOW in the series branch has two capacitors that can be strapped in for additional high frequency correction. In the shunt element following the series arm the upper potentiometer is designated FLAT and sets the mid-frequency gain through the repeater. It also is involved in correcting the cable response characteristics. The potentiometer across the *LC* mesh is designated HIGH. The HIGH potentiometer determines the amount of the high frequency peaking that occurs in the equalizer. The two coupling capacitors following the shunt element are strapped to provide low frequency correction. There is interaction between the controls as would be expected in this equalizer configuration. However, a step-by-step procedure for repeater alignment results in a fast and accurate alignment.

The amplifier consists of three common-emitter stages. The first stage has both local and loop feedback. In the second two stages the first transistor, an emitter follower, drives a common emitter transistor with feedback tieing the collector of the first transistor to the emitter of the second transistor. This is called a compound and is treated as a single transistor. The compound results in the

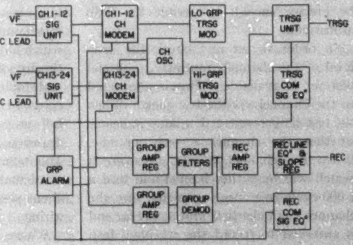


Fig. 1. Block diagram of 81-A system terminal

Paper 61-775, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 20, 1961; made available for printing May 25, 1961.

E. A. Gilmore and C. G. Griffith are both with Lenkurt Electric Company, Inc., San Carlos, Calif.

	WEST TO EAST			EAST TO WEST		
CHAN	MOD WEST TERM	LINE	DEMOD EAST TERM	MOD EAST TERM	LINE	DEMOD WEST TERM
1	388.5	178.5	199.5	388.5	185.5	192.5
2	402.5	164.5	213.5	402.5	171.5	206.5
3	416.5	150.5	227.5	416.5	157.5	220.5
4	430.5	136.5	241.5	430.5	143.5	234.5
5	444.5	122.5	255.5	444.5	129.5	248.5
6	458.5	108.5	269.5	458.5	115.5	262.5
7	472.5	94.5	283.5	472.5	101.5	276.5
8	486.5	80.5	297.5	486.5	87.5	290.5
9	500.5	66.5	311.5	500.5	73.5	304.5
10	514.5	52.5	325.5	514.5	59.5	318.5
11	528.5	38.5	339.5	528.5	45.5	332.5
12	542.5	24.5	353.5	542.5	31.5	346.5
13	388.5	346.5	346.5	388.5	353.5	353.5
14	402.5	332.5	332.5	402.5	339.5	339.5
15	416.5	318.5	318.5	416.5	325.5	325.5
16	430.5	304.5	304.5	430.5	311.5	311.5
17	444.5	290.5	290.5	444.5	297.5	297.5
18	458.5	276.5	276.5	458.5	283.5	283.5
19	472.5	262.5	262.5	472.5	269.5	269.5
20	486.5	248.5	248.5	486.5	255.5	255.5
21	500.5	234.5	234.5	500.5	241.5	241.5
22	514.5	220.5	220.5	514.5	227.5	227.5
23	528.5	206.5	206.5	528.5	213.5	213.5
24	542.5	192.5	192.5	542.5	199.5	199.5
SIG	381.5	381.5	381.5	388.5	388.5	388.5

	WEST TERM	EAST TERM
LOW GRP OSC XMIT	567.0	574.0
HIGH GRP OSC XMIT	735.0	742.0
LOW GRP OSC REC	378.0	378.0

Fig. 2 (left). Frequency allocation of 81-A system

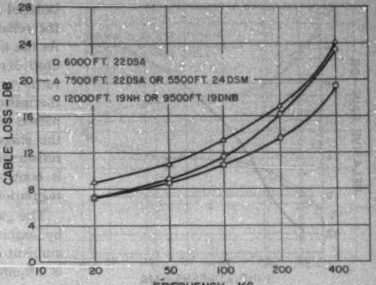


Fig. 3 (right). Typical cable losses

high-frequency characteristics of the emitter-follower and the power capabilities of the output transistor.[1] Bridge feedback is taken from the collector of the compound and returned to the emitter of the first stage. This stabilizes the output impedance. The output transformer matches the amplifier to the line. There is approximately 20 db of loop feedback and 15 db of local feedback in the first stage and the compound. The gain of the amplifier is approximately 31 db. However, the minimum loss of the equalizer at 403 kc is approximately 6 db resulting in a maximum repeater gain of 25 db.

The distortion characteristics of the amplifier are illustrated in Fig. 6 where the two-tone test and break point were used. Particular attention has been given the 2A-B products which result in intelligible intermodulation. The 2A-B products also add on a voltage basis rather than power. This results in 6-db degradation per repeater rather than 3 db per repeater. As 12 repeaters can be used in tandem in one direction of transmission this distortion characteristic is extremely important.

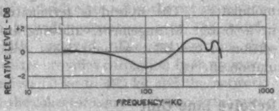


Fig. 4. Frequency response of 13 miles of 22-gage cable

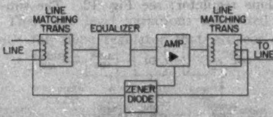


Fig. 5. Block diagram of nonregulating repeater

The repeaters are installed without any power or lightning surge protection of the carbon block, carbide, or gap type. The repeaters are physically well isolated from the cable sheath or ground and are allowed to float with the surge. The input transistor is protected from voltages developed as a result of minor imbalances in a cable pair by the equalizer, and the output transistor is protected by the resistance divider that furnishes the bridge feedback.

The repeaters were examined for lightning and surge current sensitivity by two methods. The first method applies a transverse voltage to both the input and output of the repeater by discharging a 800-microfarad capacitor in series with 10 ohms charged to 450 volts into both the input and output of the repeater. The polarity of the discharge was applied in both directions assuring that neither positive or negative surges would affect it. If the repeaters are used on a cable pair with 40-db longitudinal imbalance, then the longitudinal surge required to produce the equivalent situation is approximately 300 amperes.

The second method applied a 200 volt surge from an 800-microfarad capacitor and 10 ohms into the input and output transformers to produce a longitudinal surge through the repeater. The surge current is approximately 20 amperes. This establishes the ability of the Zener diode and filter capacitor to withstand longitudinal surges.

The use of carbons or any discharge blocks installed as protection is not recommended. Breakdown protection devices will cause extremely high transverse currents as the protectors from each side of the line to ground will not arc over at the same time. The pairs utilized in 81-A in a cable will tend to have the maximum voltage limited to the protection voltage applied to the other pairs in the cable.

The regulating repeater has an additional stage, block diagram Fig. 7, inserted between the equalizer and the amplifier used in the nonregulating repeater. In the emitter of this stage, in addition to the local feedback, there is an unbypassed thermistor. The thermistor power is derived from a resistor in series with the zener diode furnishing the power for the repeater unit. The thermistor is

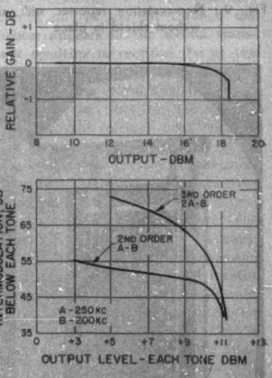


Fig. 6. Distortion characteristics of repeaters

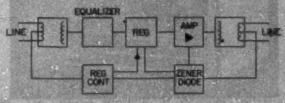


Fig. 7. Block diagram of regulating repeaters

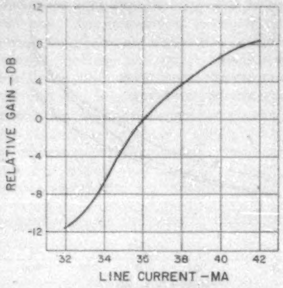


Fig. 8. Regulation response of regulating repeater

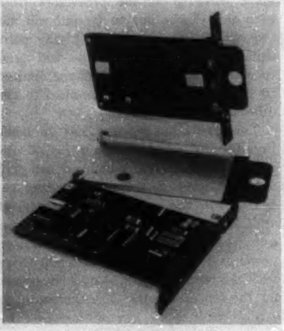


Fig. 9. Nonregulating repeater with and without shield in place

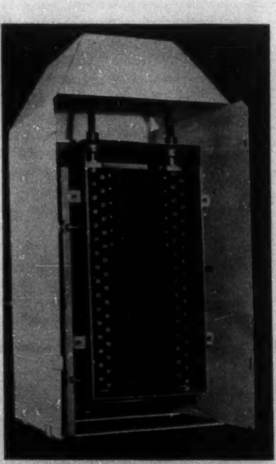


Fig. 10. Housing with repeaters and load coil

isolated for alternating current from the resistor by a 60-millihenry inductor. As the line current is varied, the voltage drop across the series resistor varies in turn, varying the current flow through the thermistor. This changes the unbypassed emitter resistance, and the gain of this stage is inversely proportional to this resistance. Thus the gain of the repeater is controlled by the line current. The regulation response is shown in Fig. 8.

The gain of the regulation is enhanced by ambient variations, that is, at high ambient temperature less current increase is required because the thermistor has been heated externally. The converse is true at low ambient temperatures. This provides additional external temperature correction.

The distortion requirements, frequency response, and lightning protection for this repeater are the same as those for the nonregulating repeater.

The high level repeater consists of two compounds, four transistors, in place of the three transistor arrangement in the nonregulating repeater. The equalizer is identical to that of the nonregulating repeater. It utilizes a silicon transistor as the final output stage and 10-db higher output level for a given input and the same distortion requirements.

In the repeaters, the use of the equalizer preceding the amplifier results in a low level hitting the first stage of the amplifier. The biasing on the first stage of the amplifier and the transistor were chosen to reduce the noise generated in the first stage. The channel noise as a result of the individual repeaters is nominally 15 decibels adjusted for the nonregulating and 18 decibels adjusted for the regulating and high level repeaters.

A repeater with and without the shield in place is shown in Fig. 9. The shield isolates the repeaters when installed in a housing and acts as a heat sink for the output transistor and the zener diode.

The repeaters are housed in a pressure tight housing with conventional cable terminals. The housing is tested at 15 pounds per square inch. The repeater housing is a hot dipped galvanized container suitable for use in manhole or aerial installation. As the cable terminals are dammed, pipe flanges are furnished in each end of the housing to allow installation of a gas line from the cable for pressurizing the housing and purging. The housing will accommodate 16 repeaters or 8 systems. The installed repeaters and load coils are shown in Fig. 10. To ease the cable splicing instructions, repeaters are furnished in two directions, East-West and West-East. This allows one cable stub to go to the East office and the second cable stub to be directed to the West office.

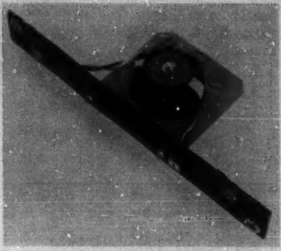


Fig. 11. Photograph of load coil

All repeaters are designed to operate in ambient temperatures of −40 to 150 degrees Fahrenheit. For aerial installations a sunshade is furnished to reduce the effect of solar radiation on the housing.

An 88-millihenry toll grade loading coil is furnished that has the same mounting facilities as the repeater so that a housing with its terminals may be completely spliced in while anticipating future growth. That is, a single system with its two repeaters may be installed in the housing, and the balance of the pairs may be utilized with the load coils as voice-frequency (v-f) pairs. As growth is encountered the load coils can be replaced with repeaters. A photograph of the load coil is shown in Fig. 11.

Transmit Line Unit

The transmit line unit of the carrier combines the two 12-channels groups and the pilot for transmission over the cable pair. The two 12-channel groups are combined in a hybrid with high trans-hybrid loss to prevent cross modulation between the two groups in the group modulators. The hybrid is terminated in a bridge arrangement, allowing insertion of the pilot without cross modulation effects.

Receive Line Unit

The receive line unit consists of a simplex transformer, an equalizer, and a slope regulator; see Fig. 12. The sim-

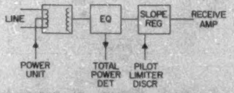


Fig. 12. Block diagram of receive line unit

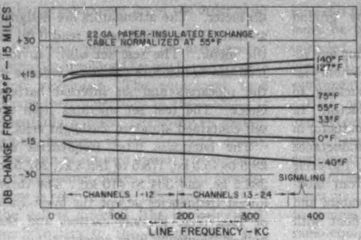


Fig. 13 (left). Temperature characteristic of cable

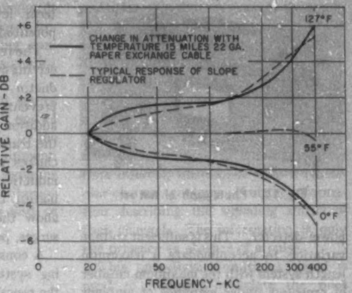


Fig. 14 (right). Slope regulator response

plex transformer provides a connection that energizes the repeater string and an impedance transformation. The equalizer compensates for the section of cable adjacent to the office. The slope regulator corrects for changes in the cable loss characteristics that occur with changing cable temperature. The temperature loss characteristics of cable are illustrated in Fig. 13.

Slope regulation is accomplished by varying the termination of the bridge tee networks in the emitter and collector of the compound amplifier. The bridge tee presents an impedance at the input that varies with frequency as the termination is varied. When the network is terminated in its characteristic impedance, the input impedance is uniform with frequency. The bridge tees terminate in thermistors whose impedance is controlled by the collector and emitter current of the output transistor of the compound. The slope regulator response for different currents is illustrated in Fig. 14. The collector and emitter currents of the output transistor are controlled by the rectified pilot from the PIL/LIMITER DISC (Fig. 12). The rectified pilot is connected to the base of the first transistor of the compound. This allows the compound to act as a d-c amplifier and slope regulator.

Two bridge tee networks are necessary to provide correction over the bandwidth involved. The network in the emitter provides correction from approximately 50 kc to 400 kc and the network in the collector corrects the hook between 20 kc and 50 kc.

Terminal Regulation

There are level variations associated with the slope regulator that are corrected by the receive amplifier unit. The receive amplifier unit, Fig. 15, consists of a total power (24-carrier) regulator, temperature compensation, and an output amplifier. The total power regulator is a two-stage amplifier with a thermistor in the feedback network. Any variations in the total power present at the output cause the thermistor impedance to change so that the amplifier gain maintains constant power output.

The feedback thermistor is sensitive to ambient variations, so temperature compensation is necessary. The compensation is an attenuator with one element containing another thermistor.

The group amplifiers are similar to the receive amplifier; they regulate the 12-channel groups.

Residual slope variations that appear at the output of the receive line amplifier are reduced by a factor of four. The receive amplifier reduces the variations by 50%, then each of the group amplifiers reduce the remaining variations by 50% because of the total power regulation.

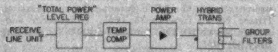


Fig. 15. Block diagram of receive amplifier

Repeater Regulation

The input to the slope regulator portion of the receive line unit is held constant. A portion of the output of the equalizer in the receive line unit is connected to the total-power-detector. This unit amplifies, rectifies the signal and compares it with a fixed voltage. The difference drives a d-c amplifier. This controls the repeater line voltage. As the input power to the receive line unit decreases, the voltage fed to the repeater string increases, increasing the power feed current. The regulating repeaters respond, increasing their output resulting in the total power being held constant at the input of the slope regulator.

The total-power-detector amplifier consists of two sections. The first section is a three-stage feedback amplifier followed by a voltage doubling rectifier. The second section consists of a silicon diode biased "on" as a voltage reference and a d-c amplifier connected as an emitter follower. The emitter follower drives the transistor voltage regulating circuit in the terminal power unit. This is illustrated in Fig. 16.

The terminal power unit is an inverter driving several transformers with tapped windings and the voltage regulator. The voltage regulator consists of two center-tapped transformers separated by a transistor in the center leg and rectifying diodes in the outer legs. The square wave appearing across the first transformer is blocked from the second transformer by the diodes when the transistor is turned off. This results in no output from the regulator. When the transistor is turned on the current through the transistor appears in the second transformer resulting in rectified d-c at the output. The current or voltage is controlled by the conduction of the transistor. The regulator furnishes 0 to 180 volts to be added to the received fixed voltages present.

The inverter powers the transformers and rectifiers furnishing the transmitting and receiving current. Strapping is provided for various system lengths. The maximum voltage available for transmitting is 144 volts in 36-volt increments and the maximum voltage available for the receiving is 156 volts in 12 volt increments. The 0 to 180 volts of the voltage regulator is normally adjusted to 90 volts and is adjusted by a potentiometer in the total-

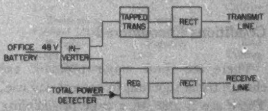


Fig. 16. Block diagram of terminal power unit

Fig. 17. Photograph of test set

power-detector. This is sufficient voltage variation to accommodate a maximum length system with the maximum number of repeaters under extremes of ambient temperatures and ground potentials caused by streetcars, voltage transients, etc.

The return for the repeater power can either be a v-f pair or earth ground. More than one system can use a v-f pair as a return. It is desirable to use a v-f pair to remove the possibility of short-term level bumps caused by ground potential variation.

The returns for the transmitting and receiving currents are through a 75-henry inductor. The currents are bucking to prevent saturation of the inductor. In addition to this is a 10-henry inductor in the transmit lead. The high-inductance chokes prevent any 60-, 120-, 180-cycle induction from power lines from modulating the output of the repeaters and allow the repeater string to float with surges, power or lightning.

A constant-current circuit is used during system lineup. This circuit allows the repeater string current to be adjusted to and held at 36 milliamperes. This prevents level variations in the regulating repeaters.

A terminal and repeater test set is used for alignment of the common equipment and repeaters. The test set consists of four filters, level correcting attenuators, an amplifier and an expanded scale, ±3-db meter. The attenuators are designed so that all test adjustments read 0 db on the meter. The test set also provides facilities to measure the zener voltage in the repeaters and an internal battery check. The test set operates from six self-contained standard flashlight cells.

The bandpass filter frequencies are 38.5 to 45.5 kc, 178.5 to 185.5 kc, 267.5 to 283.5 kc, and 364 to 410 kc. The three lowest frequencies accept the West-East or East-West channel carriers and the highest frequency filter accepts either direction of the pilot and is wide enough to accommodate the sidebands. The test set is shown in Fig. 17.

The use of the test set is such that untrained personnel can align the system with a minimum of instruction.

Reference

1. A New Compound Transistor, J. A. Stewart. *Proceedings*, Southwestern Conference, Institute of Radio Engineers, Austin, Tex., Mar. 1958, pp. 16–18.

The 81-A Exchange Trunk Carrier System—Signaling

E. F. TUCK
NONMEMBER AIEE

DONALD K. MELVIN
ASSOCIATE MEMBER AIEE

THIS PAPER describes the unique time-division-multiplex supervisory and signaling system employed in the 81-A Exchange Trunk Carrier, the general features of which are covered in a companion paper by M. E. Ferguson and M. C. Harp in this issue. In the design of carrier systems, a problem requiring almost as much effort as that of the transmission of conversations is that of the transmission of signaling information. In fact, considering the wealth of options and varieties of signaling encountered in exchange carrier service, one laboratory wag has described exchange carrier equipment as "a collection of signaling gear with incidental talking channels." And, indeed, in exchange carrier design, the variety of signaling requirements encountered is almost without end.

Conditions Encountered

There are three major types of exchange trunk signaling in present day use. These are E and M, a type of signaling used largely for toll and special exchange applications; loop dial, which is used between step by step exchanges and to some extent between crossbar exchanges; and revertive, an older type of signaling used between panel offices, between panel offices and certain types of crossbar offices, and in some cases between crossbar offices.

E and M signaling (the meaning of these two letters, if they ever had specific meaning, has long been lost) was originally developed for use in carrier systems, and is the least demanding from the standpoint of carrier system design. The signaling and voice paths are totally separate. The battery applied to an M lead at one end of a system, causes battery or ground to appear on a corresponding E lead at the distant terminal. In this system, the burden of translating subscriber dial pulses or other signaling information to a form compatible with the carrier system is placed upon accessory equipment at the terminals.

By the use of these accessory trunk repeaters E and M type signaling may be adapted to almost any type of more specialized signaling within the pulse rate range of the carrier equipment. The additional cost of trunk repeaters, however, adds substantially to the cost of the carrier channel.

In loop-dial signaling arrangements, changes in the d-c condition of the talking circuit are used to indicate various signaling conditions on the path. Table I indicates several methods of loop signaling and the d-c loop conditions existing at various times during the completion of a call. Since a carrier system is usually capable of transmitting only on-off (binary) information, the application of this type of signaling to E and M type carrier signaling equipment requires a considerable amount of conversion equipment.

The conditions existing on the subscriber loops at each end of the path are predictable only within limits. The total series resistance in the loop may vary from 1,400 ohms to as little as 100 ohms, there may be d-c leaks in parallel with the loops, and considerable dial pulse distortion may be encountered from parallel capacitance. Detection equipment applied to the line should not cause imbalance to ground, since such imbalances may result in audible noise in the subscriber sets. Typical longitudinal balances which are maintained in modern

---

Paper 61-776, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted March 20, 1961; made available for printing May 23, 1961.

E. F. Tuck and Donald K. Melvin are both with Lenkurt Electric Company, Inc., San Carlos, Calif.

exchange equipment are of the order of 60 decibels.

Revertive pulsing is an arrangement originally developed for use between panel type exchanges. The equipment that makes the connections in these offices is essentially a huge, motor-driven switch, which because of its size and design, must run at its own speed. In signaling between panel offices, the usual arrangement is reversed: the office toward which the call is directed (the incoming end) sends pulses to the calling office (the outgoing end) indicating the position of the incoming switch. These pulses appear as a series of short circuits of the interoffice trunk. When a specific number of these pulses has been received (related to the called party's number) the outgoing office responds with a momentary open circuit, ordering the incoming switch to stop at that point.

Since this stop pulse must be received at the incoming switch before it has moved to the next step, the round-trip delay involved in the signaling path becomes extremely important in revertive applications. The maximum tolerable round-trip delay in revertive interoffice trunks is about 5 milliseconds (msec). This is roughly ten times as fast as the operating time of ordinary relays; hence more sophisticated techniques are required to apply this type of signaling to carrier channels.

To complicate matters further, reverse battery is applied to the trunk at various points in the revertive pulsing cycle. This represents a third kind of information which must be transmitted in some manner through the carrier equipment.

Historically, carrier system signaling has been accomplished by tones in or near the voice band of the channel. These tones are shifted in frequency or keyed on and off to indicate various signaling states. If these tones are out of band, that is, within the spectrum assigned to the channel involved but outside that part used for voice communications, they are sometimes used for such secondary applications as channel regulation. The use of out-of-band tones for signaling places severe requirements on the channel filters, since these tones cause interference if they are allowed to leak into the voice band.

With in-band signaling, in which tones are applied within the voice band, filter requirements may be relaxed but the signaling equipment becomes much more complex. Obviously, no tones may be present when the subscribers are connected to the channel; equipment to switch the voice paths on and off during signaling intervals must be provided. Also, guards must be provided to prevent voice frequencies, that are present during conversation, from operating the signaling equipment.

Both systems require a signal source and detection equipment at each end of each channel. This adds greatly to the complexity of the channelizing equipment. In addition, the presence of the signaling tones may constitute a large contribution to system loading on repeaters and on radio equipment.

Theory

To avoid the problems inherent in the use of signaling tones in conventional or carrier signaling, and to meet the necessity of producing extremely short round-trip delays for revertive applications, a radically new approach has been used in the Type 81-A Exchange Carrier. In this approach a time-division-multiplex system is used, in which each channel is sampled in turn. The resulting time-division-multiplex signal is transmitted to the far end of the system and, in turn, each channel at the far end receives information regarding the d-c condition at the near end.

In this approach, as in time-division-multiplex systems of all kinds, most of the complexity is in the common equipment. Channel units become very simple indeed. This method of signaling has the additional advantage of making a negligible contribution to system loading, because it utilizes a pilot which would be present in any case. This system also enjoys the well-recognized frequency-modulation advantages. The approach used allows the use of on-off, or binary-type circuits at almost all points. Though a relatively large number of active components is utilized, the circuitry is highly repetitive and therefore relatively simple to service.

A block diagram of the time-division multiplex signaling system is shown in Fig. 1. In principle, the operation of the system is quite simple: a master clock oscillator at one terminal of the system, through counters and sampling gates, samples each channel in turn to determine its d-c condition (signaling status). The result of this sampling is a time-division-multiplex signal in which a discrete time interval, identifiable with a particular channel, contains all of the information describing the signaling status of that channel. The information is only the presence or absence of a pulse.

This time-division-multiplex signal is used to frequency-modulate the pilot signal transmitted along with the speech channels in the carrier equipment. When this pilot is demodulated at the distant terminal, it yields the previously described time-division-multiplex signal. This signal is then disassembled and routed to the various channel signaling units at the distant terminal by a process exactly the reverse of that by which it was assembled. In order to time this process correctly, a clock oscillator at the distant terminal is made to run in step with the master clock at the near terminal.

The system described here, when considered a black box, will respond to a negative voltage applied to a sampling lead at one terminal, with a negative-going pulse on the corresponding output lead at the other terminal. The transmission time through the black box varies from a few microseconds to 2 msec, depending upon the point in the scanning cycle at which the input signal is applied.

The translation from the loop condition or other signaling condition to the black box input-output conditions is handled by channel signaling units associated with the individual speech channels. These channel signaling units are of various types, depending upon the particular type of signaling information to be transmitted.

Table I. Signaling State of Typical Loop Signaling Systems

	Calling End		Called End		
Type of Supervision	Seizure (Connect) Signal	Disconnect Signal	On-Hook	Off-Hook	Battery Source
High-Low	Closed Loop	Open Loop	High Resistance	Low Resistance	Calling End
Low-High	Closed Loop	Open Loop	Low Resistance	High Resistance	Calling End
Reverse Battery	Closed Loop	Open Loop	Normal Line Polarity	Reverse Line Polarity	Called End
Reverse High-Low	Closed Loop Followed By Battery Reversal	Battery Returns To Normal Polarity, Followed By Open Loop	High Resistance	Low Resistance	Calling End
Wet-Dry	Closed Loop	Open Loop	Battery And Ground Connected To Line	Battery And Ground Removed	Called End

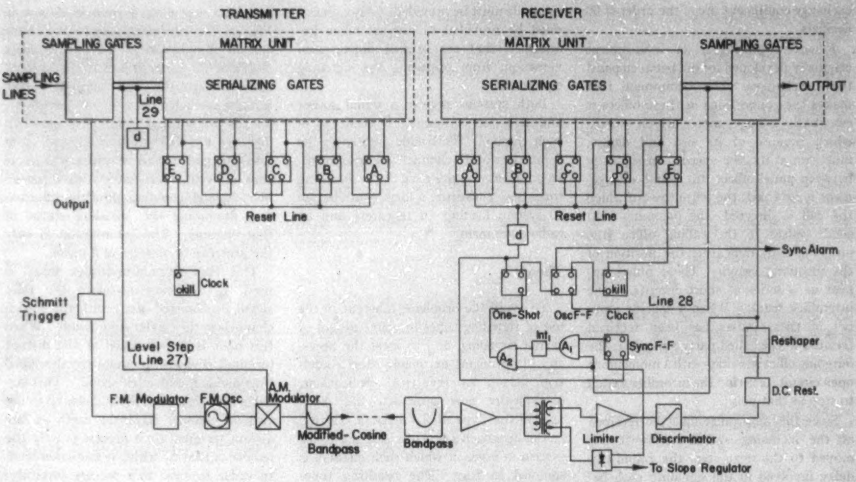

TRANSMITTER
RECEIVER
SAMPLING GATES
MATRIX UNIT
SAMPLING LINES
SERIALIZING GATES
Line 29
d
E
D
C
B
A
Reset Line
Output
Clock
kill
Schmitt Trigger
Level Step (Line 27)
F.M. Modulator
F.M.Osc
A.M. Modulator
Modified-Cosine Bandpass
Bandpass
Limiter
Discriminator
To Slope Regulator
MATRIX UNIT
SERIALIZING GATES
SAMPLING GATES
OUTPUT
A
B
C
D
E
Reset Line
d
Sync Alarm
One-Shot
OscF-F
Clock
kill
Line 28
Int1
A1
A2
Sync F-F
Reshaper
D.C. Rest.

Fig. 1. Block diagram of the time-division-multiplex signaling system

Description

The following paragraphs describe in some detail the design and operation of the time-division-multiplex signaling system. In the process of this description frequent reference will be made to Fig. 1, the block diagram, and Fig. 2, which is a timing diagram showing the relationships between the voltages at various points, and at various times. This illustration covers two complete 2-msec frames of signaling information. It shows the condition of all 24 channels, the alarm unit, and synchronizing pulses which will be described.

The transmitting clock oscillator that supplies the master timing pulses for the

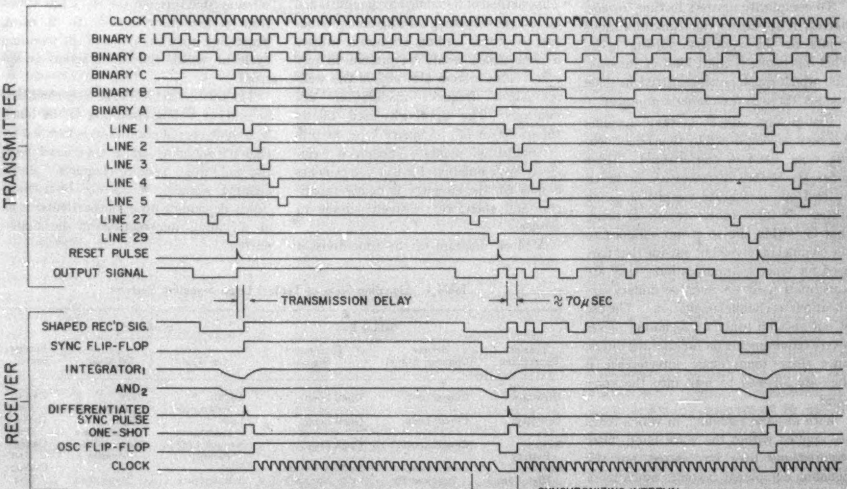

TRANSMITTER
CLOCK
BINARY E
BINARY D
BINARY C
BINARY B
BINARY A
LINE 1
LINE 2
LINE 3
LINE 4
LINE 5
LINE 27
LINE 29
RESET PULSE
OUTPUT SIGNAL
TRANSMISSION DELAY
≈ 70μ SEC
RECEIVER
SHAPED REC'D SIG.
SYNC FLIP-FLOP
INTEGRATOR1
AND2
DIFFERENTIATED SYNC PULSE
ONE-SHOT
OSC FLIP-FLOP
CLOCK
SYNCHRONIZING INTERVAL

Fig. 2. Timing diagram of the time-division-multiplex signaling system

The first frame assumes all channels idle; the second indicates some signaling in progress

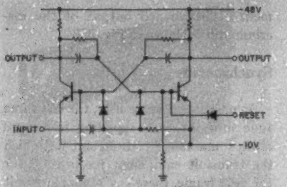


Fig. 3. Schematic diagram of typical binary counter stage

whole signaling system is essentially a free running multivibrator, stabilized by a tuned circuit. Its output waveform is labeled Clock in Fig. 2. This oscillator produces a positive going edge approximately every 70 μsec. This edge is used to trigger the E binary first, which in turn, on every second oscillator pulse, triggers the D binary, which in turn, on every fourth oscillator pulse, triggers the C binary, etc. It is evident that if each of these binaries is examined after each of the oscillator pulses there will be 32 discrete states of the outputs of all of these binaries before the sequence begins to repeat itself. A schematic diagram of a typical binary is shown in Fig. 3.

Only 30 bits of information are used in this system; therefore, the binaries are reset to the initial zero state after the twenty-ninth oscillator cycle. Thus, the set-up, which is depicted in Fig. 2, begins to repeat itself after every thirtieth transmitter clock cycle.

The output voltage waveforms of each of the five binary stages composing the binary counter chain are shown in Fig. 2. There are actually two outputs from each binary, one the inverse of the other. (These inverse outputs are not shown.) The other output of the E binary, for instance, is similar to the E output shown, but is 180 degrees out of phase. It is referred to as $\bar{E}$ (not E).

The binary code output of the binary counter chain is not useful as it stands; it must be converted to a serial form to provide a series of discrete sampling pulses. This operation is accomplished in a set of diode serializing gates.

These gates consist of a matrix of diodes, with 10 binary inputs and 31 serial outputs. Diode gates are arranged within this matrix in such a manner that all output lines are held to a low negative voltage until the binary counter output corresponding to the number of that line appears. The line is then allowed to become more negative if a negative voltage is applied externally. This action may be understood by reference to Fig. 4, a schematic diagram of the matrix unit.

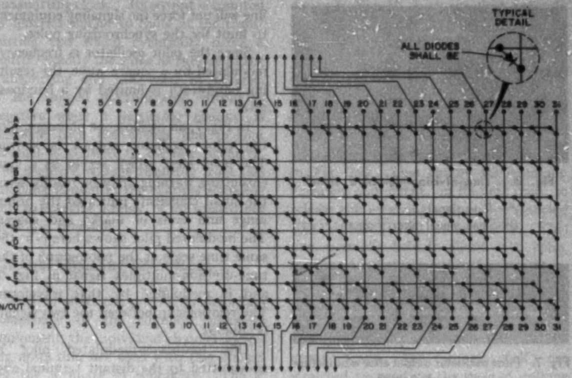


Fig. 4. Schematic diagram of matrix unit

Ignoring the effect of the line labeled IN/OUT for the moment, it is evident that line 23 may become negative if, and only if, input lines A, $\bar{B}$, C, D, and E are negative. In Fig. 2 it is indicated that this situation occurs only during the time interval labeled 23, (as mentioned previously, line $\bar{B}$ is negative when line B is positive).

In this manner, each of these output lines is released in turn, beginning with line 1. It is during this released interval that the channel signaling unit, whose number corresponds with that of the line, is sampled. If the line, while released, has a negative voltage applied to it within the channel signaling unit, the line labeled IN/OUT will also be carried negatively. The diodes that connect each sampling line to the IN/OUT line are the sampling gates.

At the end of a complete frame (the period during which each channel is sampled once), the IN/OUT line will have indicated the condition of each channel in turn by its voltage during the period when the channel was sampled. The voltage appearing on this line is then the raw output of the time-division-multiplex signaling system.

The schematic complexity of the matrix unit is slightly misleading. The unit itself, physically, is quite small, as Fig. 5 shows. The unit actually contains 332 diodes, since two diodes are used in series at each junction for added reliability. The diodes used are approximately 1/8-inch diameter by 0.035-inch thick. The mechanical arrangement may best be described by analogy to a Swiss cheese sandwich, in which the input and output lines form the bread and the diodes are placed in holes in an insulating Swiss cheese. The internal layout of diodes is the same as that of the schematic, except that the whole assembly is folded back upon itself at line 15.

After all 24 channels are sampled, the system alarm unit is sampled to determine whether a group alarm exists. Lines 27, 28, and 29 are permanently connected (through resistors) to negative battery; therefore they place three permanent marks on the output at the end of each frame. These marks are used for synchronization purposes, as will be explained.

The raw time-division-multiplex output of the sampling gates as it appears on the IN/OUT matrix line is shown in Fig. 6(A). The effects of the matrix diode capacitances are quite pronounced here, because the source impedance of the channel signaling units is quite high. This waveform is shaped at this point by a Schmitt trigger, the output of which is shown in Fig. 6(B).

This shaped waveform is used to frequency-modulate the pilot oscillator.

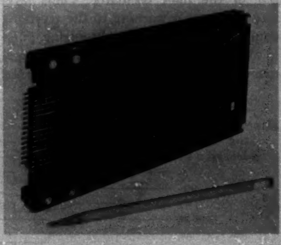


Fig. 5. Photograph of matrix unit

Fig. 6. Time-division-multiplex signal

A—Raw signal
B—Same signal after shaping

Fig. 7. Pilot oscillator output after modulation and band limiting

Fig. 8. Time-division-multiplex signal

A—After demodulation
B—Same signal after shaping

The pilot used in 81-A Carrier is on a nominal frequency of 381.5 or 388.5 kc depending upon the direction of transmission. This is the center frequency of the pilot. The actual frequency at which the pilot may be found is either 14.4 kc above or below this frequency, depending on whether a mark or space is being transmitted. To assure reliable synchronization, and to avoid synchronization on the arrangements of marked channels that might resemble the synchronizing pulse, an amplitude level step is added to the output of the pilot oscillator. This level step serves as a warning to the receiving equipment that the synchronizing pulse is about to be received. This arrangement assures that a momentary hit on the line will not force the signaling equipment to hunt for the synchronizing pulse.

Since the pilot oscillator is frequency-modulated by a square wave, the resulting bandwidth is limited by a bandpass filter.

In Fig. 7, an oscilloscope photograph of the pilot oscillator output is shown. The level step, which serves as a synchronizing warning, may be readily seen; the small transients indicated the beginnings and ends of marks and spaces. The modulated pilot shown here is at the same time scale as the photograph of the time-division-multiplex signal in Fig. 6, and is modulated by that waveform. The relationship between the two signals is obvious.

This frequency-modulated pilot is transmitted to the distant terminal and amplified. A portion of the pilot is rectified after amplification and used to provide a reference voltage for a slope regulator which maintains constant drop levels in the carrier system.

Any amplitude information which has been modulated on the pilot is recovered along with the d-c level. The warning level step is separated at this point. The remainder of the pilot signal is applied to a two-stage limiter that removes residual amplitude and noise components. The output of the limiter is applied to a discriminator, the output of which is shown in Fig. 8(A).

It may be seen that some distortion has occurred in transmission. The particular signal exhibited in Fig. 8(A) had been transmitted through an artificial line approximating actual operating conditions of the carrier system. The rounding of the pulses is caused by the bandwidth limitation imposed by the bandpass filter in the pilot oscillator at the transmitting end and by in-route phase versus frequency nonlinearities encountered in the cable.

In spite of this distortion, all information originally contained in the signal may be recovered by shaping the signal as was done in the transmitting equipment. The shaped output of the discriminator is shown in Fig. 8(B).

Synchronism

As mentioned previously, the last three time units of each signaling frame are always marks. At the point at which the transmit equipment resets itself for the next frame, there is a 1-sampling unit space. These four pulses, the three marks and the space, constitute the synchronizing pulse. These are indicated in the timing diagram, Fig. 2.

The detailed mechanism of synchronization is as follows: the level step, as detected and shaped, is used to set a flip-flop in the synchronizing unit; see Fig. 9, an enlarged timing diagram of the synchronizing interval, and Fig. 10, a block diagram of the synchronizing unit. This SYNC flip-flop, when set, opens a diode gate, which allows the shaped discriminator output to be applied to the one-shot multivibrator. The one-shot multivibrator is tripped by the first positive-going edge as shown in Fig. 9. This positive-going edge is the edge that occurs between the last mark and the space of the synchronizing pulse. It is to this edge that all receiving-end timing is referred. This one-shot multivibrator is adjusted to exactly the length of this last space; therefore, when the one-shot multivibrator resets itself, the information that corresponds to the sampling of Channel 1 is about to be received.

When the one-shot multivibrator is set it causes the SYNC flip-flop to be reset, closing the gate to its input. When the one-shot multivibrator resets it sets the oscillator flip-flop. The oscillator flip-flop controls the clock oscillator at the receiving end. It is at this point that a similarity to telegraph practice may be seen. The clock oscillator will easily run in phase with the transmit clock oscillator for a few cycles, in this case 28 cycles. At this point the clock oscillator turns itself off by resetting the oscillator flip-flop. When the next synchronizing

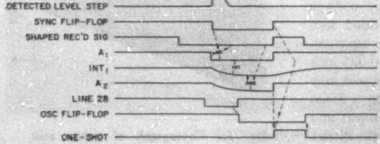


Fig. 9 (left). Enlarged portion of Fig. 2, showing the synchronizing interval in detail

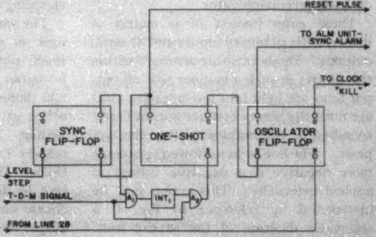


Fig. 10 (right). Block diagram of synchronizing unit

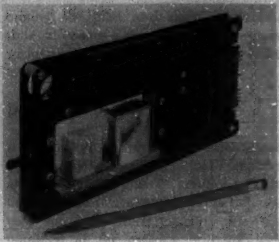


Fig. 11. Photograph of E and M signaling unit

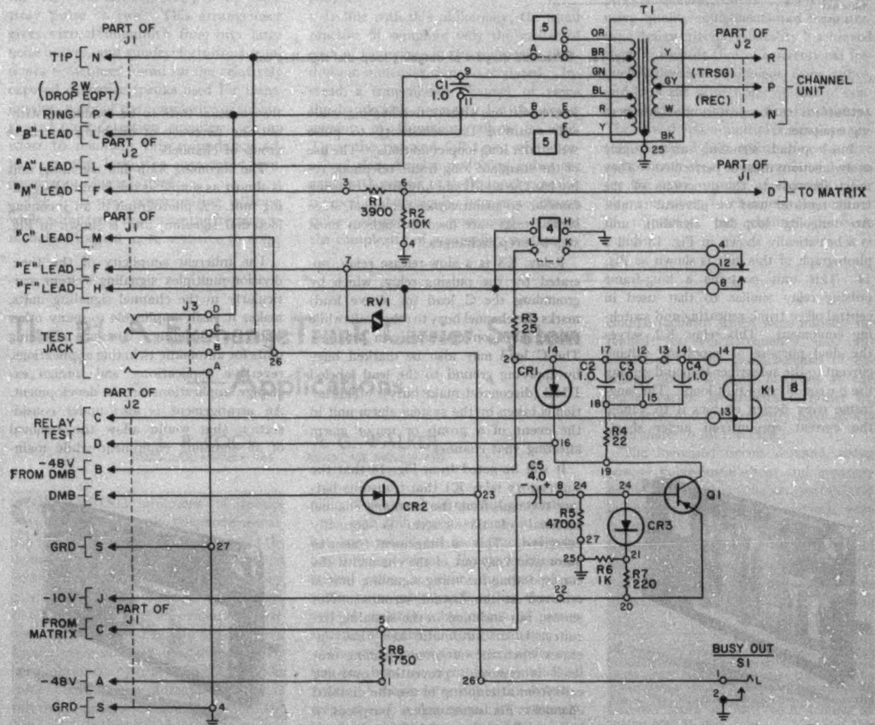
pulse arrives, the clock oscillator at the receiving end is turned on again. In this manner the clocks at each end of the system are kept in phase and the channels are sampled at the correct intervals.

The receiving clock is identical to the transmitting clock. Its output is counted by binaries similar to those at the transmitting end which eventually drive a matrix, exactly the same as the one at the transmitting end but used in reverse. If the shaped discriminator output is negative when a particular channel is sampled, then the output line to that channel is allowed to become negative. The negative voltage resulting from this sampling is used in the channel signaling unit to energize a relay, or for some other signaling function.

Channel Signaling Equipment

As if to offset the previously described complexity of the common signaling equipment, the individual channel signaling units that are used with this system are the ultimate in simplicity. The E and M signaling unit, shown in the photograph of Fig. 11 and schematically drawn in Fig. 12, is an excellent example of this simplicity.

In E and M applications, similar signaling units are used at both terminals of the same channel. If the M lead (transmit lead) is connected to a negative battery at one terminal, a negative voltage is applied through resistor R1 to the channel sampling line (labeled TO MATRIX in Fig. 12). This causes the common signaling equipment, on its next scan, to release the receiving output line (labeled FROM MATRIX) momentarily. This line, normally held to about −10 volts, is allowed to become suddenly more negative under the influence of R8. The resulting negative pulse is amplified by Q1 and charges capacitors C2, 3, and 4, operating relay K1 after two or three successive scans. The contacts of K1 are strapped to produce the desired voltage on the E lead (receive lead) which is employed by the user of the equipment to



PART OF J1
TIP N
2W (DROP EQPT)
RING P
"B" LEAD L
PART OF J2
"A" LEAD L
"M" LEAD F
PART OF J1
"C" LEAD M
"E" LEAD F
"F" LEAD H
J3
TEST JACK
D
C
B
A
26
PART OF J2
RELAY TEST C
D
−48V FROM DMB B
DMB E
GRD S
−10V J
PART OF J1
FROM MATRIX C
−48V A
GRD S
T1
5
C
D
A
OR
BR
GN
BL
R
Y
B
E
F
5
BK
25
Y
GY
W
(TRSG)
(REC)
PART OF J2
R
P
N
CHANNEL UNIT
C1 1.0
9
11
PART OF J1
D
TO MATRIX
3
R1 3900
6
R2 10K
4
4
J
H
L
K
7
RV1
R3 25
2
CR1
14
17
12
13
14
C2 1.0
C3 1.0
C4 1.0
18
15
14
13
K1
8
R4 22
16
19
C5 4.0
8
24
24
Q1
R5 4700
CR3
27
25
R6 1K
21
R7 220
20
22
23
CR2
5
R8 1750
1
26
4
4
12
8
BUSY OUT S1
1
2
L

Fig. 12. Schematic diagram of E and M signaling unit

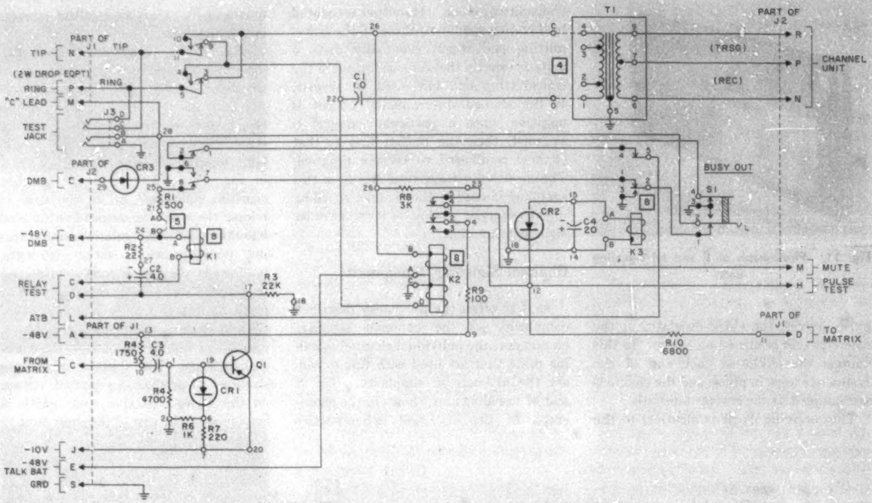

PART OF J1
TIP
N
(2W DROP EQPT)
RING
P
"C" LEAD
M
TEST JACK
J3
PART OF J2
DMB
C
CR3
R1 500
-48V DMB
B
R2 22
C2 40
K1
RELAY TEST
C
D
ATB
L
PART OF J1
-48V
A
R4 1750
C3 4.0
FROM MATRIX
C
R4 4700
CR1
Q1
R6 1K
R7 220
R3 22K
-10V
J
-48V TALK BAT
E
GRD
S
C1 1.0
T1
4
R8 3K
K2
R9 100
CR2
C4 20
K3
BUSY OUT
S1
(TRSG)
(REC)
PART OF J2
R
P
N
CHANNEL UNIT
M
MUTE
H
PULSE TEST
PART OF J1
D
TO MATRIX
R10 6800
5
8

Fig. 13. Schematic diagram of outgoing loop-dial signaling unit

actuate his trunk repeaters or other signaling equipment.

In loop-dial signaling units, many more functions must be performed. They must incorporate the functions of the trunk repeater used on physical trunks. An outgoing loop-dial signaling unit is schematically shown in Fig. 13 and a photograph of this unit is shown as Fig. 14. This unit contains a long-frame pulsing relay similar to that used in central office trunk repeating and switching equipment. This relay, K2, serves the dual purpose of supplying talking current to the subscriber loop and sensing the d-c condition of that loop. The long-frame relay design enables it to handle the current encountered under short-loop conditions without overheating, while still retaining the sensitivity to pulse well under long-loop conditions. The use of the standard long-frame telephone relay has the additional advantage of being familiar to maintenance personnel, since similar relays are used throughout most step-by-step exchanges.

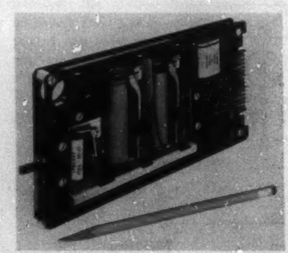


Fig. 14. Outgoing loop-dial channel signaling unit

Relay K3 is a slow-release relay, operated by the pulsing relay, which by grounding the C lead (or sleeve lead) marks the channel busy to other calls while pulsing and conversation is in progress. The C lead may also be marked busy by applying ground to the lead labeled DMB (disconnect-make-busy). This action is taken by the system alarm unit in the event of a group or major alarm affecting that channel.

It will be noted from Fig. 13 that the supervisory relay K1 that transmits battery reversals from the incoming channel terminal of the system is normally energized. This arrangement assures automatic busy-out of the channel if the corresponding incoming signaling unit is removed at the distant terminal. Not shown, but included in the signaling circuitry, is an automatic key that operates when the outgoing signaling unit itself is removed, preventing outgoing calls from attempting to use the disabled channel. For maintenance purposes, a busy-out key is accessible from the front of the signaling unit allowing pulsing and transmission tests to be conducted without the necessity of disabling an entire group of channels.

The incoming loop-dial signaling unit is almost as simple as the E and M signaling unit. A photograph of an incoming loop-dial signaling unit is shown in Fig. 15.

The inherent simplicity of the time-division-multiplex signaling system, particularly in the channel signaling units, makes it easily applicable to many other forms of signaling. Channel signaling units for automatic ticketing applications, revertive applications, and foreign exchange applications are in development. An arrangement is also under consideration that would allow the removal of all signaling equipment while main-

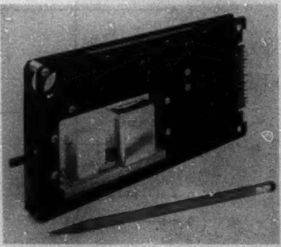


Fig. 15. Incoming loop-dial signaling unit

taining the pilot and slope regulation; thus, affording a saving in cost to users wishing to employ multitone or other forms of in-band signaling.

Design Philosophy

In this system, certain deviations were deliberately made from normal computer and data transmission practice. It is usually true in communications practice that simplicity and reliability may be obtained at the expense of bandwidth and redundancy. Since the signaling information to be transmitted is quite slow by data transmission standards (a fast telephone dial generates about 12 pulses per second, a normal one about 10) an average pulse is sampled about 50 times during its lifetime. The channel signaling units that employ relays are accordingly designed so that if one or two scans are missing during the on-time of a pulse, the relay will not fall out; the relay will likewise not pull up on a stray pulse or two. This arrangement gives virtual immunity from hits, large noise spikes, and sundry fortuitous transients sometimes found on the relatively exposed exchange trunks used for transmission, without the complexity of sophisticated error-detection circuitry. In an effort to maintain this redundancy advantage the level-step-augmented synchronizing scheme is used.

The level-step synchronizing scheme, while admittedly rendering the frequency-modulated signal more sensitive to noise, insures that synchronization will immediately occur at the correct point upon resumption of service. Since even the worst noise spike will obliterate only a frame or two, with an assured synchronizing point the over-all effect at the channel signaling units is nil. The over-all effect of noise, therefore, is reduced for the type of information being transmitted.

It should be noted that an optimally designed binary frequency-modulation data transmission system will operate with a negligible error rate up to an 8-decibel signal-to-noise ratio. Obviously, this is an unusable level for voice communications. Since the signaling system provides alarm functions as well as signaling functions, it is desirable to have a noise failure occur in the signaling system at roughly the same point at which the talking channels become unusable. This has the simple humanitarian function of not allowing a subscriber to complete (and pay for) a call when conversation is impossible.

In line with this philosophy, the usual practice of sampling only the center of each of the pulses in the composite time-division-multiplex signal is obviated. Instead, a transmission channel of twice the bandwidth necessary for the transmission of Gaussian pulses is used. Thus, the rise and fall times of the received time-division-multiplex pulses are small with respect to the channel sampling interval, and ambiguity from nonoptimum pulse shapes is minimized without the complexity of a sampling system.

In the past, time-division-multiplex schemes have been approached with justifiable caution because of the disadvantage of a component failure (particularly a tube) in the relatively complex common equipment imposed on the integrity of the communications service. Due to the combined advances in transistor technology, computer techniques, and reliability engineering, this position is no longer tenable. Switching circuits are characteristically relatively insensitive to any changes in device parameters, and, therefore, they are almost completely unaffected by considerations of device linearity.

Generally, at the speeds considered here, a switching circuit does not perform more or less well; it either works or it doesn't. Therefore, equipment life is primarily dependent on castastrophic device failures. It is the catastrophic failure aspect of transistor life that is best understood, and responds best to in-process quality control. Therefore, by using quality components and computer-type design criteria, reliability is achieved that far exceeds that of conventional frequency-division equipment using tubes, and is of the same order as that of conventional equipment using transistors. The time-division-multiplex signaling system described here utilizes fewer active components than a conventional system for the same number of channels. For larger numbers of channels, the number of transistors per channel decreases rapidly.

The 81-A Exchange Trunk Carrier System —Applications

L. R. COOL
ASSOCIATE MEMBER AIEE

R. G. WALKER
ASSOCIATE MEMBER AIEE

CONTINUING programs of service improvement by the telephone operating companies have sharply increased the demand for telephone circuits in recent years. For the long-haul applications, carrier systems have been used extensively to satiate this demand. However, for short trunks of 15 miles or less, circuit expansion has almost exclusively been obtained by the expansion of the cable plant. Since a major portion of the requirements for new short-haul circuits exists in the large metropolitan areas, the installation of new cable is often difficult as a result of conduit congestion.

As with toll circuit expansion, carrier has been considered as a practical solution to this problem. Except for the most critical cases, however, the cost of installing new cable has been less than the cost of expansion by means of carrier systems designed for toll circuit use.

The requirement for a new carrier system has been apparent for a number of years. A major deterrent to the development of such a system was its extremely low cost objective. In addition, the system would normally be applied to exchange facilities which were initially intended for voice-frequency use. Information on the characteristics of exchange and toll-grade facilities at the higher carrier frequencies is rather incomplete, and only recently have efforts been made to fill the gaps.[1]

The increased circuit demand, rising cost of cable installation, and improvements in the state of the art, particularly in the area of solid-state devices, were factors that weighed heavily in favor of the design of a carrier system that could compete favorably with new cable installations. Economic prove-in curves for 81-A exchange trunk carrier system

---

Paper 61-777, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18-23, 1961. Manuscript submitted March 20, 1961; made available for printing May 23, 1961.

L. R. Cool and R. G. Walker are both with Lenkurt Electric Company, Inc., San Carlos, Calif.

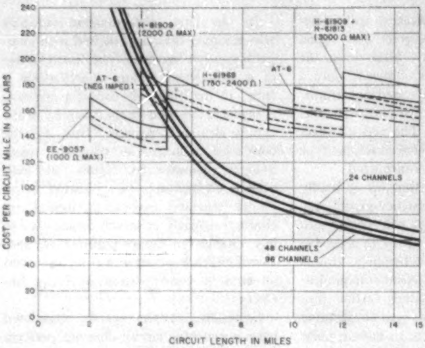


Fig. 1. Comparison of installed costs of exchange trunks

First cost prove-in, 22-gage paper-insulated lead-covered

Solid line—51 pair, 25 used

Broken line—101 pair, 50 used

Irregularly broken line—202 pair, 100 used

compared with several commonly used types of cable are shown in Fig. 1. Cable costs in this illustration include total installation and material based on average value of 125% of material and include negative impedance repeaters or trunk repeaters suitable for the distance indicated. Carrier costs include all equipment, suitable repeaters, and installation based on 140% of equipment cost.

To satisfy the low-cost objective, the requirements for the system were carefully studied in terms of market potential and design feasibility. Based on these studies, the development program for short-haul carrier system was initiated. The 81-A system, which was recently introduced, is the result of this development program. Specific design features of the 81-A system are described in detail in companion papers in this issue. In this paper, factors which are important to the application of the system are considered.

Area of Application

The 81-A system was developed to permit orderly and economical expansion of short-haul trunks in the exchange plant. Each system provides a maximum of 24 channels on two cable pairs. In contrast with carrier designed for toll use, exchange trunk carrier must, where possible, reduce prove-in costs and operate directly into the office switching equipment. For example in loop signaling, the signaling units of the 81-A system provide the necessary trunk circuitry to permit operation between the outgoing selector banks and the incoming trunk selectors as shown in Fig. 2.

Although most applications would involve only interlocal or extended area service, it was anticipated that carrier derived trunks may also be used as toll connecting links from class 5 offices. Accordingly, the 81-A system was designed to permit operation at via-net-loss +2 decibels (db), and the two-wire voice-frequency hybrids were designed to meet the echo requirement of 18-db average and singing point of 10-db average. It was also felt necessary to provide as a strappable option, either 600 or 900 ohms drop impedance for toll and local switching requirements.

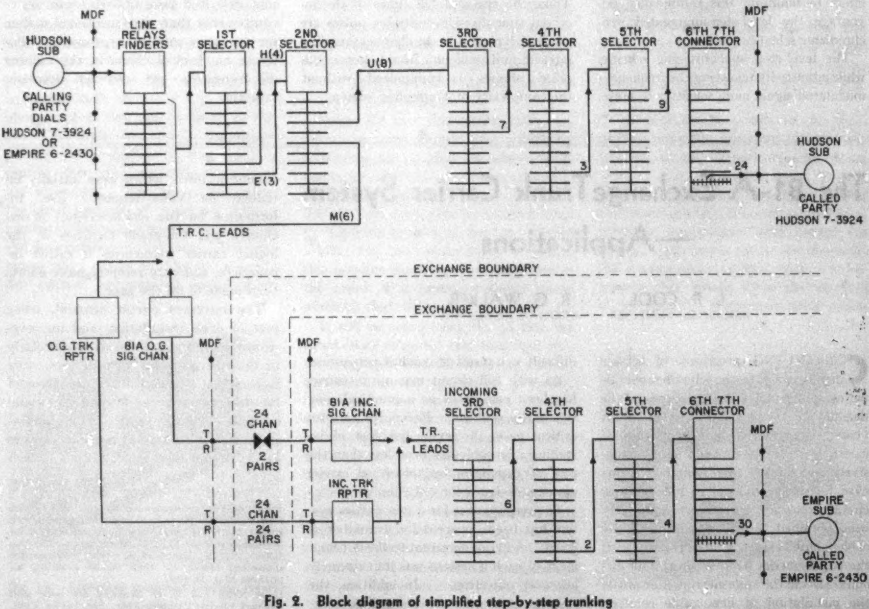

MDF
HUDSON SUB
CALLING PARTY DIALS
HUDSON 7-3924 OR EMPIRE 6-2430
LINE RELAYS FINDERS
1ST SELECTOR
2ND SELECTOR
H(4)
E(3)
U(8)
M(6)
3RD SELECTOR
4TH SELECTOR
5TH SELECTOR
6TH 7TH CONNECTOR
7
3
9
24
MDF
HUDSON SUB
CALLED PARTY HUDSON 7-3924
T.R.C. LEADS
EXCHANGE BOUNDARY
EXCHANGE BOUNDARY
O.G. TRK RPTR
81A O.G. SIG. CHAN
MDF
MDF
T
R
24 CHAN
2 PAIRS
24 CHAN
24 PAIRS
81A INC. SIG. CHAN
INC. TRK RPTR
T.R. LEADS
INCOMING 3RD SELECTOR
4TH SELECTOR
5TH SELECTOR
6TH 7TH CONNECTOR
6
2
4
30
MDF
EMPIRE SUB
CALLED PARTY EMPIRE 6-2430

Fig. 2. Block diagram of simplified step-by-step trunking

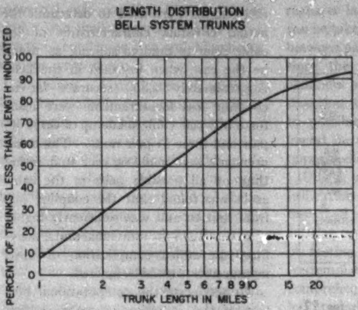


Fig. 3. Trunk distribution in the Bell System

In the independent telephone companies approximately 94% of the trunks are less than 15 miles in length

In the early study phases of the project, it became apparent that restrictions must necessarily be placed on both the type of facility and the length over which the system could be expected to operate, if the cost objectives and transmission performance objectives were to be maintained. Because of the stringent cost objectives, elaborate techniques for frequency synchronization, noise and crosstalk suppression, regulation, and slope correction must necessarily be avoided. Selection of the modulation method and the frequency range were made to obtain the maximum advantages that were consistent with minimum cost.[2]

In direct contrast with the necessity of restricting system applications was the requirement for high-volume production as an important means of obtaining lower costs. This demanded a careful weighing of the cable installation costs against design objectives, trunk lengths as a function of the percentage of trunks, and the performance characteristics of the facilities most likely to be encountered. Since 22-gage 0.083-microfarad-per-mile cable comprises over 40% of the existing mileage in the exchange plant, the basic design parameters were established for this cable type.

A prove-in cost objective was tentatively set for 5 miles. In actual practice, systems have proved-in over distances of less than 2 miles in length. The ultimate system length is primarily dependent upon the transmission performance of the cable and the transmission advantages that can economically be incorporated in the design of the system. Taking all factors into account, it was concluded that the design objective be set for a maximum system length of 15 miles on 22-gage exchange cable, and that the design be optimized for an 8-mile system length. Referring to Fig. 3, the 3- to 15-mile range includes approximately 45% of the existing trunks.[3]

Although 22-gage cable would be most commonly encountered, the quantities of 19-, 24-, and 26-gage high-capacity exchange cable and 19-gage low-capacity cable were significant and could not be ignored in the design. Therefore, sufficient equalization range was included in the terminals and repeaters to accommodate applications to all types of cable.[4]

Thus, the 81-A system may readily be applied to the various types of cable normally found in the exchange plant. However, with the smaller gage cable, the maximum system length is somewhat less than 15 miles due to transmission and repeater power feed considerations.

Transmission Considerations

In the 81-A system, the approximate frequency spectrum of 20 to 400 kc is used in both directions of transmission. Because of this broad frequency range, detailed transmission calculations could become quite involved, particularly if slope variations had to be determined by each of the various types of cable making up the facility.

In the 81-A system, this problem was greatly simplified by establishing a standard repeater spacing and providing adjustable compensation for deviations from the standard interval. The standard repeater spacing was chosen as 6,000 feet in 22-gage high-capacity cable. This is the normal loading interval for *H* loaded cable, and at 400 kc, this represents a flat loss of 20 db.

For system engineering acceptable variations from the standard interval are given in terms of length for the different types of cable as shown in Table I. The ranges are sufficient for most cases and can often be used in selecting the best repeater locations.

A further simplification in system engineering was achieved by providing three different types of repeaters.[4] A typical layout is shown in Fig. 4. Since the 81-A is a four-wire interlaced system, each repeater is a one-way device in that it provides amplification for only one direction of transmission. The nonregulating repeater (NR) provides a maximum gain of 24 db at 400 kc. The gain-frequency characteristic is controlled by an adjustable equalizer. Once set, the gain-frequency characteristic is relatively constant over the normal expected ambient temperature range. This repeater is most commonly used because of the relatively small loss changes that occur in a repeater section. A regulating repeater (R), with a gain-frequency characteristic similar to that of the NR repeater, provides about ±5 db of flat loss correction at 150 kc. Spacing of the R repeaters depends primarily upon the cable temperature range expected. The high-level (HL) repeater provides an additional 10-db gain over that available from the NR repeater. The HL repeater is most commonly located at the half-section loading point adjacent to an office.

The use of the different types of repeaters was standardized for normal applications. The HL repeaters normally work into an office and the increased level provides additional margin against impulse noise in the cable. The R repeaters would precede an HL repeater. For systems requiring additional flat loss correction, spacing of additional R repeaters would be made so that each

Table I. Recommended Repeater Section Lengths and Variations in Length That Can Be Accommodated by Built-In Adjustment

		19-Gage Cable						22-Gage Cable, 0.082μfd/M			24-Gage Cable, 0.084μfd/M			26-Gage Cable, 0.079μfd/M		
		0.062μfd/M*			0.084μfd/M											
		R	Min	Max	R	Min	Max	R	Min	Max	R	Min	Max	R	Min	Max
Repeater section lengths in feet	End section	4700	2400	6600	3800	1900	5400	3000	1500	4100	2200	1100	3100	1800	1000	2600
	Intermediate section	9000	3800	12000	7700	3100	9200	6000	2350	7000	4500	1800	5300	3600	1500	4400

* Microfarad per mile.

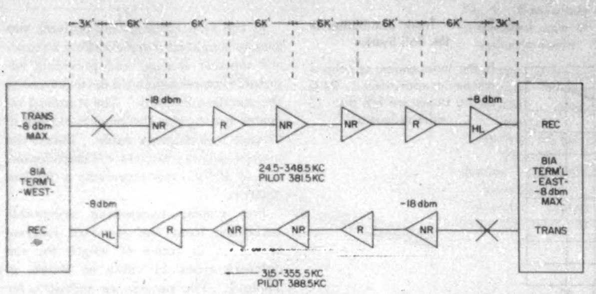


Fig. 4. A system layout example showing repeater location on standard intervals for 22-gage exchange cable

R—Regulating NR—Nonregulating HL—High-level

(3,000-foot sections at terminal ends with high levels were used to overcome effects of office-induced impulse noise. The levels shown are for individual channel carrier power values.)

repeater would compensate for about the same magnitude of flat loss change. The remaining repeaters would consist of the NR type.

The availability of the three types of repeaters has added greatly to the flexibility in system applications. As an example, for unavoidably long repeater section lengths, which may occur at underwater crossings, the additional gain of a HL repeater may be used to advantage. In contrast to this, extremely short end sections are occasionally encountered in the telephone plant. In some cases, even the reduced output level of an HL repeater is too high to be easily accommodated by the receiving terminal equipment. For such applications, an R repeater can be used. This avoids overloading of the receiving terminal equipment while still maintaining adequate receiving transmission levels.

Cable Crosstalk

As only very limited data were available for exchange type cables covering the carrier frequency range up to 350 kc, it was decided that the necessary data to confirm or disprove our theories concerning crosstalk in exchange type cables should be obtained. In addition to the frequency-attenuation temperature characteristics, which must be compensated for, the effect of crosstalk coupling losses on system performance must be evaluated.

Pair selection would be the choice of the most widely separated pairs which would normally have the higher far-end crosstalk coupling loss, however, straight splicing has not been a normal practice, particularly in the older types of cable which are most likely to be encountered in existing plant facilities. With these older types of cable, improved performance at voice frequencies has been obtained by random splicing, and it was expected that random splicing would most often be encountered. Under this condition pair selection would be meaningless unless transmission measurements were made for each pair selected. However, some advantage can be gained in the newer unit-type cable where random splicing is restricted to each unit. Pairs can be selected from different units for carrier applications, and the full advantage of separation between units can be realized.

For this analysis measurements were conducted on several 1,500-foot reel lengths of 303-pair 22-gage paper-insulated lead-covered cable. Standard reel lengths were used to determine the actual crosstalk characteristics of the cable prior to random splicing as would be the case when installed in the field. Approximately 7,350 separate far-end coupling loss measurements were made for a 303-pair cable made up of three 50-pair and three 51-pair units. The measurements were made on each unit rather than on all possible pairs in the cable as it was found that the coupling loss from unit to unit was sufficiently higher than the rms of the individual units, therefore, the added complication of these measurements could be ignored.

To accelerate the computational portion of this investigation, each reading was punched on a separate IBM (International Business Machines) card, and the cards for each unit of the cable were identified and stacked in separate decks. Initially, the statistical distribution of each unit was obtained by sorting and the rms coupling and sigma were then determined. The average values of far-end coupling loss are shown in Fig. 5. The spread between units was within ±3 db of the average for the six units. Curve 1 is an average of 7,350 pair-to-pair couplings in six units of a 22-gage 303-pair paper-insulated lead-covered cable 1,500 feet long. The remaining curves are the computed result of random splicing of two, five, and 15 sections of the cable.

Of primary concern was the far-end crosstalk coupling loss resulting from a random combination of a number of cable sections. For two sections, the cards of two different cable units were individually shuffled to give a random distribution. Next the cards were alternately selected from the two decks by the IBM machine (i.e., random spliced). The individual coupling losses were converted to power values, totaled, and re-

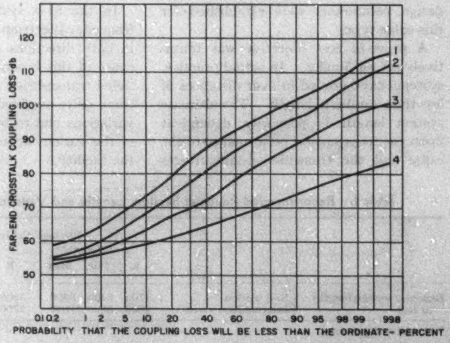
Fig. 5. Far-end crosstalk coupling loss data

1—303 pair of paper-insulated lead-covered 22-gage cable, one 1,500-foot section

2—2 random spliced sections

3—5 random spliced sections

4—15 random spliced sections



converted into coupling loss on a new card. In a similar manner, the resultant effects of combining five and 15 splicing sections were obtained.

To insure a high degree of accuracy from the available data, random splicing was simulated 896 times for hypothetical cables made up of two or five splicing sections and 902 times for the cable made up of 15 splicing sections. For each hypothetical cable, the number of occurrences for each value of far-end crosstalk coupling loss was recorded. Beginning with the lowest coupling loss a running total was made from which the probability of occurrence was computed. These results are also shown in the probability graph, Fig. 5. As was to be expected, there was a good possibility that, as the number of splicing sections increased, a section with a low coupling loss would be connected with a number of sections having relatively high loss. The net effect is shown rather dramatically in the 99% region of the graph. An increase of from two to five sections causes a decrease of about 10 db, and from five to 15 sections there is an additional decrease of about 18 db. Fortunately, the coupling losses in this region were sufficiently high so that the resulting decreases that occurred were of minor consequence.

Of greater importance is the probability that a number of sections, each with a very low coupling loss, would be spliced together. While this possibility exists, the convergence of the curves for different numbers of splicing sections indicates that this probability is remote. The effect of increasing the number of random splices appears to be that of improving the uniformity, and with only a gradual lowering of the coupling loss for the worst cases. This latter fact indicates that the decrease in far-end crosstalk coupling loss is not a function of 10 log N, where N is an integral multiple of a unit length, for individual coupling losses in random spliced cable. However, the rms coupling of a cable unit may be expected to decrease on a 10 log N basis.

For stagger-twist cable, it is expected that 50% or more of the pairs will be usable in most applications without pair selection. This is an increase of about 600% in the number of circuits that can be obtained from the cable.

Neglecting the subjective crosstalk advantage designed into the system, the odds on the usability of pairs can be improved by selection.[3] In most cases, random splicing has been confined to units within the cable. Where this has been done, pairs may be selected from different units and spliced into the repeater housing. Thus, full advantage can be taken of the increased coupling loss between pairs from different units. Also, since the 81-A system uses interlaced channels and a separate pair for each direction of transmission, it will often be possible to use those pairs with the lowest far-end crosstalk coupling loss for opposite directions of transmission.

Unless the maximum possible pair utilization is anticipated, the previously described selection methods should prove satisfactory, and they require a minimum of testing. If maximum utilization is contemplated on random spliced cable, comprehensive far-end crosstalk coupling loss tests would be required. In this case, for new installations of paper-insulated lead-covered cable, measurements should be performed on each cable section. Pairs should then be spliced to obtain the highest coupling losses at 350 kc for those pairs assigned to carrier. On the new polyethylene cables, straight splicing will permit direct selection of pairs most suitable for carrier applications.[1]

Equipment Features

In keeping with the over-all low equipment cost requirements, equipment and transmission engineering, installation, and maintenance should also be low. In addition, a new application area for carrier was to be invaded, therefore, personnel who would be installing and maintaining the equipment would be relatively untrained in carrier techniques.

Mechanically, the equipment would necessarily differ from that normally found in the exchange offices, but some of the advantages have been taken in the techniques used for wiring and cabling. Rear access would normally be provided, and for this reason, external as well as intrarack wiring could be made directly to connectors at the rear of the rack. This not only afforded an equipment cost reduction over toll carrier equipment that provides terminal blocks, but it followed established techniques of wiring and gave the appearance of belonging in the office for which it was designed.

New manufacturing methods were fully exploited to achieve minimum cost with maximum reliability and ease of maintenance. Card-type mechanics were used, utilizing a stitched-wiring technique rather than printed circuit boards. This technique has a number of advantages; among these are ease of repair and flexibility in circuit design changes. The circuit board cards plug into standard sized shelves that are equipped with connectors. Shelf and external wiring complete the connection to other circuits and/or to the h-f (high-frequency) line and drop circuits. Two photographs are shown in Fig. 6 that illustrate the use of unit extendors and test equipment. The Terminal and Repeater Test Set being used are shown in Fig. 6(A).[4] Fig. 6(B) illustrates the use of the voice-frequency test set.[2] Test jacks on the extension units facilitate lineup measurements and adjustments.

A—Terminal and repeater test set with units on extendors

B—Voice-frequency test set connected to a channel under test

Fig. 6. Use of test equipment

An important point in circuit board design is to maintain a functional grouping of related circuits on the same card. In many cases, this would require that cards be of varying sizes. As an alternative, where a larger card size is required, two or more cards may be combined to provide the necessary mounting space and the physical size of the cards remain the same. In this way a standard shelf size can be maintained and automatic manufacturing methods may be set up more easily. In the 81-A system, only two different card sizes are used. The smaller card mounts in a shelf that requires $3^1/_2$ inches on a standard 19-inch equipment rack. The larger card mounts in a shelf that requires $8^1/_4$ inches on a 19-inch equipment rack. Several multiple cards have been used to maintain the functional grouping of the circuitry.

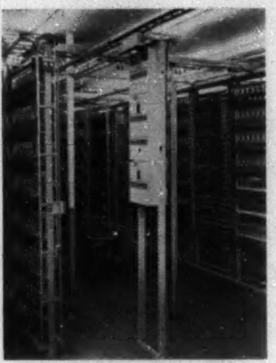


Fig. 7. Photograph of typical installation of 84 channels with 48 volt power filter at top

Testing of equipment units during production is facilitated by maintaining the functional circuit grouping. More important, equipment routines and troubleshooting techniques are greatly simplified. These factors are important adjuncts to any carrier system, but, because maintenance will be carried out by relatively untrained personnel, this makes ease of troubleshooting a more important requirement for exchange trunk carrier system applications. It is expected that troubleshooting will primarily concern itself with isolation of the trouble to a particular unit and replacement of the defective unit. Repair of the defective unit would be accomplished on a "time permits" basis at a local repair center, or it may be returned to the manufacturer.

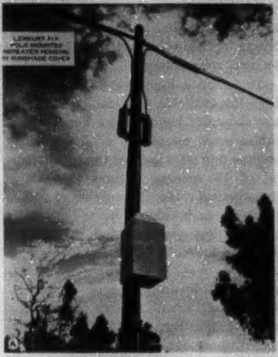
A spare carrier line facility and patch panel has been made available as an option to simplify the problem of isolating a defective repeater or bad cable pair. This panel is used, in conjunction with a separate, spare repeater string to provide an alternate transmission path, and repeater power is supplied by the panel to maintain the repeater circuits in operating condition. Patching facilities are provided for both the h-f line circuits and for the repeater power-feed circuits for each system operating over the same cable. In the event of failure of the transmission medium, a rolling patch can be made to restore service immediately. Location of the fault can then be accomplished at a more leisurely pace.

Terminals

A photograph of a typical installation is shown in Fig. 7; a photograph showing call-outs of some of the terminal equipment may be seen in Fig. 1 of the Ferguson and Harp paper, p. 601 of this issue. A type 81-A terminal is composed of three basic parts. The 48-volt power filter, which provides the necessary isolation from the office equipment operating on the common −48-volt office battery supply, and is arranged to mount on the cable rack above the terminal equipment. The carrier and power supply shelf, which provides fusing and the regulated terminal power voltages, repeater power feed source and the twelve basic channel carrier oscillators. The carrier and power supply shelf is common to four 24-channel terminal shelves. Each terminal shelf provides mounting facilities for the 24 individual plug-in signaling and channel units, and the common equipment which provides additional stages of modulation, amplification, and the necessary automatic flat loss and slope correction facilities. Maximum use has been made of the miniturization possible with solid-state devices at the present state of the art, and a complete 96-channel terminal may be mounted in one 11-foot 6-inch equipment rack.

Since the carrier will be used to provide local trunk circuits, similarity of use between the carrier and office switching circuits demanded that the signaling facilities of the carrier should be as much like that used in normal voice-frequency interoffice trunks as possible. By the use of plug-in signaling equipment, greater flexibility was possible than is usually encountered in the normal interoffice trunking facilities. For this reason, regular telephone-type relays were used, and the circuits of the loop signaling equipment, for example, were designed to perform the functions of a trunk repeater. In this way, operation between selectors could be achieved and the similarity with normal office trunking equipment was maintained. In addition, busy-out keys were included on all signaling units. This permits individual channel busy-out in the same manner as the busy-out keys used on voice-frequency trunks. Because of the plug-in feature, removal of a signaling unit would permit calls to land on an unusable trunk that would not give the calling party any indication of a busy or faulty trunk condition. For this reason, a spring contact was provided in the terminal shelf for each signaling position, thus, removal of a signaling unit caused a ground to be placed on the *C* or control lead to indicate a trunk busy condition.

Another feature of the signaling units is the butt-in or test jack for voice-frequency monitoring and test purposes. This jack is designed to accept the standard 4-conductor butt-in test plug currently in use by the local central office personnel.



A—Pole line

B—Manhole

Fig. 8. Typical repeater installation

Repeaters

Because of the high losses encountered at the carrier frequencies used in the 81-A system, repeater spacings would necessarily be short if large level differentials were to be avoided. In 22-gage exchange cable *H* loading with attendant 6,000-foot loading intervals is commonly employed. A simplification and cost reduction in the installation was anticipated if this same interval could be employed for repeater spacings. Also, the use of the system would be in congested areas, and an additional requirement was that the mounting for the repeaters should be arranged so that the repeaters could be located on existing poles or in existing manholes. Typical manhole and pole line installations are shown in Figs. 8(A) and 8(B). Repeater housings include mounting facilities for 16 one-way repeaters and may be mounted on existing poles or in manholes as shown in these photographs of typical installations. For pole-mounted housings, a sunshade is provided to reduce solar heat rise. Physical size limitations dictated a repeater design of minimum complexity, but one that would provide the necessary equalization, amplification, and, where required, regulation. The normal frequency of use of repeaters in such short systems also dictated a high degree of reliability. In addition, the physical size limitation indicated that remote power feed must necessarily be employed.

A rather unique power-feed arrangement was devised that permitted the power to be fed to the repeaters from the terminal ends only. The use of two cable pairs for the two directions of transmission meant that separate repeaters would be used for the West-East and East-West directions of transmission. Power for operation of the repeaters could be simplexed using the pair for each direction of transmission as the power-feed path for the repeaters on that pair and using a common return. The return can be either a separate cable pair or office ground. For reliability reasons, it is desirable to limit the number of systems that use the same common pair return. In this way, positive power can be applied from one terminal and negative power from the other. This permits the total voltage requirement to be split between the two terminals so that the voltage supplied from one terminal would not be excessive; this factor was considered for personnel safety.

For the 81-A system, repeater power is derived from the power unit provided for each terminal. Positive power is supplied from the transmitting end and negative power is supplied from the receiving end of the series power-feed arrangement for each direction of transmission. Control of the gain of the regulating repeaters is accomplished by changes in repeater current, and the current is controlled by the received carrier power at the terminals.[4] The receiving portion of the repeater power feed circuit, in the common power unit, includes a voltage regulator which is under control of the received carrier power.

Alarms

Automatic alarms are provided for each individual 12-channel group and are activated by a loss of total carrier power, loss of synchronization of the signaling system, or loss of battery voltage.[5] Dry contacts are provided for connection to external visual and audible office alarms. Alarm lamps mounted on the alarm unit included on each terminal shelf help to localize the equipment that is in trouble. Three alarm lamps are used, a separate lamp for each 12-channel group, where a failure occurs that is peculiar to a group, and a major alarm lamp. The major alarm occurs when the complete system is out of service. In addition, an alarm cutoff switch that silences the audible alarm is included and an associated white guard lamp. Two additional switches are provided, one for each 12-channel group, that permit busy-out on a group basis.

When an alarm indication occurs after a 10-second delay, the trunks are automatically disconnected for a period of 1 second to release the switching equipment. Following this delay, the channels are made busy by placing a ground on the control or *C* lead to the switching equipment. At the same time, information is transmitted to the distant terminal indicating that a failure has occurred. This will initiate the alarm system at the distant terminal. In this way, subscribers will be disconnected at both ends. Once the condition causing the alarm has been cleared, the alarm circuits are automatically restored to normal, and the group-busy condition is removed allowing the equipment to accept traffic again.

References

1. TRANSMISSION PROPERTIES OF POLYETHYLENE-INSULATED TELEPHONE CABLES AT VOICE AND CARRIER FREQUENCIES, **G. S. Eager, Jr., L. Jachimowicz, I. Kolodny, D. E. Robinson.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, Nov. 1959, pp. 618–40.

2. THE 81-A EXCHANGE TRUNK CARRIER SYSTEM, **M. E. Ferguson, M. Harp.** *Ibid.* (See pages 601–06 of this issue.)

3. WHERE WE ARE AND WHERE WE ARE GOING IN TELEPHONE TRANSMISSION, **H. R. Huntley.** *Ibid.*, vol. 76, Mar. 1957, pp. 54–63.

4. THE 81-A EXCHANGE TRUNK CARRIER SYSTEM—REPEATERS AND REGULATION, **E. A. Gilmore, C. G. Griffith.** *Ibid.* (See pages 606–10 of this issue.)

5. THE 81-A EXCHANGE TRUNK CARRIER SYSTEM—SIGNALING, **Donald K. Melvin, E. F. Tuck.** *Ibid.* (See pages 610–17 of this issue.)

Synthesis of Networks Simulating Impedance of Loaded Cable Facilities

R. W. DEMONTE
MEMBER AIEE

Synopsis: This tutorial paper describes a procedure for designing a network that simulates the image impedance of a loaded cable. The problem is reduced to one of synthesizing three 2-terminal impedances, two at one frequency and the third at three frequencies. In general, the return loss performance of the network against the facility is greater than 30 db (decibels) over the frequency range from 0.006 to six times the cutoff frequency, and is greater than 40 db over most of the transmitted band. With the aid of artificial cable and a loading coil, such networks can be used to simulate both the transmission and impedance characteristics of any type of loaded cable facility having any fractional length of end section.

NEEDS EXIST in the Bell System for precision 2-terminal networks to simulate the impedance characteristics of any loaded cable for test purposes. Also, 2-terminal networks are used to terminate electrically short-loaded transmission lines, and more often in hybrid circuits to balance the impedance of the transmission line.

The simple design procedure described here synthesizes impedance-simulating networks in a straightforward manner. This method is based upon the impedance or the primary constants of the cable and loading coil at frequencies located in the neighborhood of the lower edge of the desired transmitted band, the cutoff, and in the attenuated range. From these data three 2-terminal impedances are synthesized and connected in series. Both real and imaginary parts of the characteristic impedance very closely approximate that of the physical loaded cable over the entire voice-frequency

---

Paper 61-945, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, Salt Lake City, Utah, August 23–25, 1961. Manuscript submitted May 19, 1961; made available for printing June 16, 1961.

R. W. DEMONTE is with Bell Telephone Laboratories, Inc., Murray Hill, N. J.

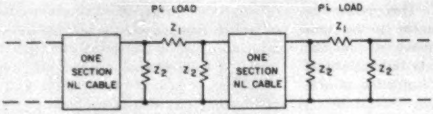


Fig. 1 (above). Line with π-network loading

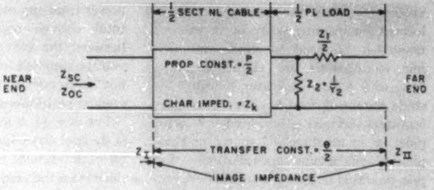


Fig. 2 (right). One-half section of π-loaded line

range starting at 20 cps (cycles per second) through the cutoff region, and in the attenuated region to 20 kc.

The use of these networks in conjunction with artificial cable[1] and loading coils permits simulation of the impedance and transmission performance of any loaded cable facility with relative precision.

Statement of the Problem

The objective is to simulate the characteristic midsection impedance of a cable for a given loading and spacing. To compute the characteristic accurately, it is necessary to have an equivalent network of the loading coil in addition to the primary cable constants. Since the loading coil is inserted in series in the line, the capacitance and conductance between windings should be taken into consideration. Furthermore the elements in the equivalent network, with the exception of the capacitance, vary with frequency as do the cable constants.

The line with π-network loading can be represented as shown in Fig. 1 where Z_1 represents the unbalanced equivalent effective series impedance of a 2-winding balanced loading coil. Z_2 represents its shunt impedance due to distributed capacitance and conductance between windings.

Now consider the 1/2-section of the π-loaded line as shown in Fig. 2. Then it can be readily shown[2] that

$$Z_{SC}=Z_K\left[\frac{1+\left(Y_2+\dfrac{2}{Z_1}\right)Z_K\tanh\dfrac{P}{2}}{\left(Y_2+\dfrac{2}{Z_1}\right)Z_K+\tanh\dfrac{P}{2}}\right] \quad (1)$$

and

$$Z_{oc}=Z_K\left[\frac{1+Y_2Z_K\tanh\dfrac{P}{2}}{Y_2Z_K+\tanh\dfrac{P}{2}}\right] \quad (2)$$

where

$$Z_K=\sqrt{\frac{R+jL\omega}{G+jC\omega}}$$

$$P=\sqrt{(R+jL\omega)(G+jC\omega)}$$

and $R,L,C,$ and G are the primary cable constants. The midsection image or characteristic impedance Z_I of the loaded cable and the transfer or propagation constant θ per section of loaded cable are computed from

$$Z_I=\sqrt{Z_{sc}Z_{oc}}=R_I-jX_I \quad (3)$$

and

$$\theta=2\tanh^{-1}\sqrt{\frac{Z_{sc}}{Z_{oc}}} \quad (4)$$

Design Procedure

The early work of Hoyt[3] and Zobel[4] suggests that a configuration made up of three 2-terminal networks connected in series can be designed to provide a good match to the image impedance of a loaded cable. A graph of the image impedance of a typical loaded cable is shown in Fig. 3.

The network to simulate this impedance will consist of three parts, namely:

$$Z_N=Z_B+Z_{S1}+Z_{S2} \quad (5)$$

where

Z_B is the basic network, Z_{S1} is the low-frequency supplementary network, and Z_{S2} is the high-frequency supplementary network. The network will have the configuration shown in Fig. 4.

The performance will be determined by the return loss against the characteristic impedance Z_I, which is

$$RL=20\log_{10}\left[\frac{Z_I+Z_N}{Z_I-Z_N}\right]\text{ in db} \quad (6)$$

Basic Network

The configuration shown in Fig. 5 for Z_B is the same as a midshunt of a double M derived low-pass filter except for two additional resistances R_1 and R_2. The coefficients of the reactive elements are based on values of $m=0.7230$ and $m'=0.4134$.

The problem here is to modify the Z_B impedance function of the filter in the vicinity of its cutoff frequency so that both components match the loaded facilities impedance components. This is illustrated in Fig. 6. Note that the cutoff frequency of the facility is defined where the absolute values of the real and imaginary impedance components are

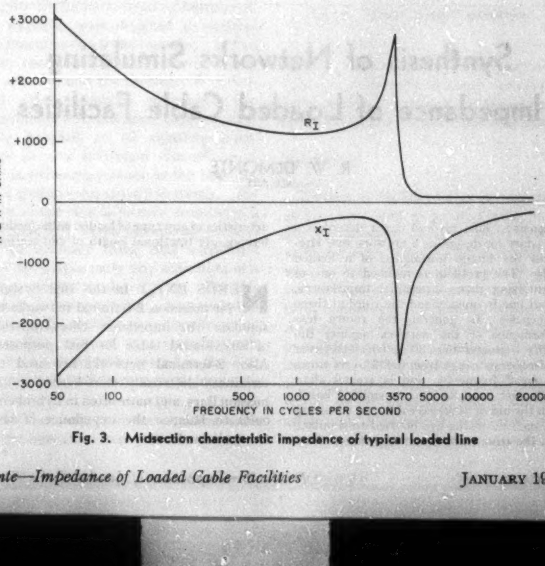


Fig. 3. Midsection characteristic impedance of typical loaded line

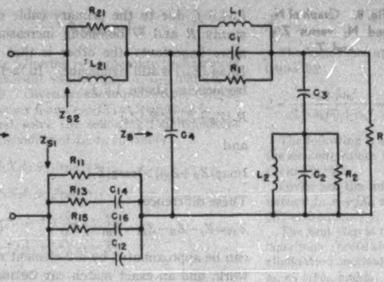


Fig. 4 (left). Schematic of network simulating loaded line impedance

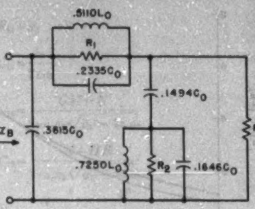


Fig. 5 (right). Element values for basic network Z_B

equal. The real and imaginary components of the driving point impedance of the filter are shown for the case where R_1 and R_2 are omitted. It was found that the addition of the two resistances R_1 and R_2 modified as desired the filter impedance function in the neighborhood of the cutoff frequency without appreciably affecting the impedance at lower or higher frequencies.

The values of the reactive elements and of the terminating resistance are determined from the primary constants of the cable and loading coil as follows:

Let

$$L_0 = L + L_1 \tag{7}$$

where L is the cable inductance per section and L_1 is the loading coil inductance per section, and let

$$C_0 = C + C_2 \tag{8}$$

where C is the capacitance of cable per section and C_2 is the capacitance between windings of the loading coil.

These constants, except possibly C_2, vary with frequency and their value should be taken at the cutoff frequency f_c.

Then

$$R = \sqrt{\frac{L_0}{C_0}} \tag{9}$$

and

$$f_c = \frac{1}{\pi\sqrt{L_0 C_0}} \tag{10}$$

from which

$$L_0 = \frac{R}{\pi f_c} \tag{11}$$

and

$$C_0 = \frac{1}{\pi f_c R} \tag{11A}$$

Since it is desirable to control the magnitudes of both impedance components of network Z_B at the cutoff frequency f_c, two resistances R_1 and R_2 are necessary and sufficient. At f_c, appropriate values of R_1 and R_2 produce an impedance Z_B with an angle of $\pi/4$ radians and therefore matches the impedance of the facility Z_I.

After normalization of element values assuming $\omega_c(=2\pi f_c)$ and R equal to unity and normalization of Z_I by dividing it by R, the schematic in Fig. 7 is obtained where

$$N_1 = \frac{R_1}{R}$$

and

$$N_2 = \frac{R_2}{R} \tag{12}$$

The expressions for determining the values of N_1 and N_2 in terms of the other network elements and Z_I' or Z_B' are rather complicated. To avoid these computations, graphs shown in Fig. 8 give the values of N_1 and N_2 for all practical values of Z_I. When $Z_I' = 3.75$, both resistances N_1 and N_2 have the value of 20.

The resistors corresponding to N_1 and N_2 should be increased to allow for the conductance introduced at antiresonance by the dissipation associated with the inductors.

This completes the determination of the element values for the basic network Z_B.

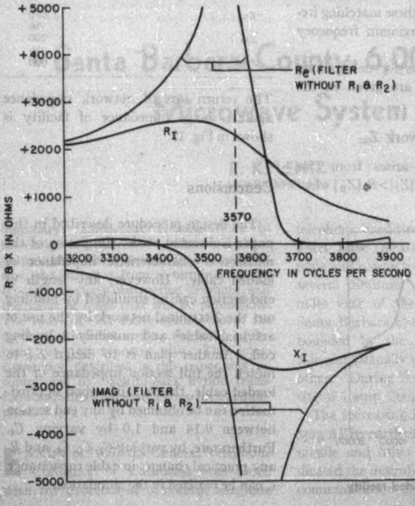


Fig. 6 (left). Enlarged graph of Fig 3 and filter around f_c

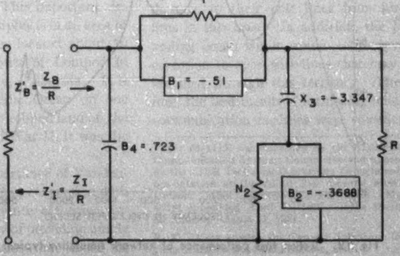


Fig. 7 (below). Element values for basic network and Z_I normalized

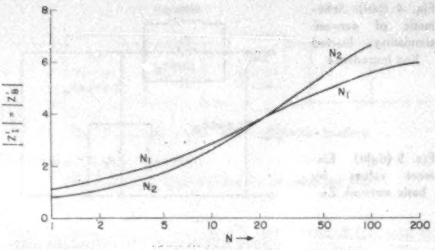


Fig. 8. Graph of N_1 and N_2 versus Z'_B and Z'_I

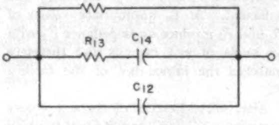


Fig. 9. Schematic of low-frequency supplementary network with four elements

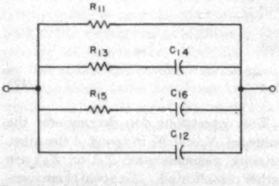


Fig. 10. Schematic of low-frequency supplementary network with six elements

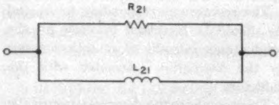


Fig. 11. Schematic of high-frequency supplementary network

Supplementary Network Z_{S1}

Z_{S1} is the low-frequency network and has impedance values equal to the difference between Z_I in equation 3 and Z_B in equation 5, Z_{S2} being considered as negligible in the frequency range below the cutoff frequency. Zobel found that when the d-c resistance of the cable is relatively small compared with R in equation 9, this impedance can be fairly accurately approximated by a simple 4-element *RC* (resistance capacitance) structure as shown in Fig. 9. However, in general, if a good impedance match is desired down to 20 cycles, then Z_{S1} should be represented by a 6-element structure as shown in Fig. 10.

The 4-element structure is intended to match the desired impedance, both real and imaginary parts, at two frequencies. The 6-element structure is intended to do the same, but with matches at three frequencies. Some judgment must be exercised in determining these matching frequencies to obtain maximum frequency coverage.

Appropriate formulas for determining these element values are given in the Appendix.

Supplementary Network Z_{S2}

The need for Z_{S2} arises from two sources; one is that $R_e[Z_I] > R_e[Z_B]$ when $f > 1.5\ f_c$ due to the primary cable constants R and G becoming increasingly more important; the other is that the Imag $[Z_{S1}]$ is still significant. In fact, at frequencies above 1.5 f_c,

$$R_e[Z_B + Z_{S1}] < R_e[Z_I]$$

and

$$\text{Imag}[Z_B + Z_{S1}] > \text{Imag}[Z_I]$$

These differences

$$Z_{S2} = Z_I - Z_B - Z_{S1} = R_{S2} + jX_{S2} \tag{13}$$

can be approximated by a 2-element network, and an exact match can be made at a single frequency, say $f = 4\ f_c$, employing the network shown in Fig. 11 where

$$R_{21} = \frac{R_{S2}^2 + X_{S2}^2}{R_{S2}}$$

and

$$L_{21} = \frac{R_{S2}^2 + X_{S2}^2}{2\pi f X_{S2}} \tag{14}$$

If the dissipation in inductor L_{21} is small, the $R_e[Z_{S2}]$ at frequencies $< f_c$ will be negligible.

Application

A simulating network has been designed based on the preceding method and on data for a loaded cable facility comprising 26-gage cable with 88-millihenry loading coils spaced at 6,000 feet. The control frequencies were selected to be

Network	Frequency, Cps
Z_B	3,570
Z_{S1}	50 200 650
Z_{S2}	8,000

The return loss of network impedance against image impedance of facility is shown in Fig. 12.

Conclusions

The design procedure described in this paper is directed at the simulation of the midsection characteristic impedance of loaded cable. However, any length of end section can be simulated by building out the 2-terminal network by the use of artificial cable[1] and possibily a loading coil. Another plan is to design Z_N to match the full section impedance of the loaded cable, then a fairly good approximation can be obtained by any end section between 0.14 and 1.0 by varying C_4. Furthermore, by varying C_1, C_2, C_3, and R, any practical changes in cable capacitance C can be realized in the simulation.

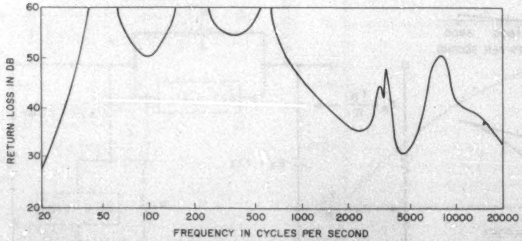


Fig. 12. Return loss performance of network simulating typical loaded facility

Appendix

The following is a procedure to determine the element values R_{11}, R_{13}, C_{12}, and C_{14} in Fig. 9. Given are the impedances r_1+jX_1 at frequency f_1 and r_2+jX_2 at frequency f_2.

First solve the following four equations for the values of a_0, a_1, b_1, and b_2:

$$a_0+f_1X_1b_1+f_1^2r_1b_2=r_1$$

$$a_0+f_2X_2b_1+f_2^2r_2b_2=r_2$$

$$a_1f_1-f_1r_1b_1+f_1^2X_1b_2=X_1$$

$$a_1f_2-f_2r_2b_1+f_2^2X_2b_2=X_2$$

which give

$$a_0=r_1-f_1X_1b_1-f_1^2r_1b_2$$

$$a_1=r_1b_1-f_1X_1b_2+\frac{X_1}{f_1}$$

$$b_1=\frac{f_1f_2(f_1X_1-f_2X_2)(r_1-r_2)+(f_1X_2-f_2X_1)(f_1^2r_1-f_2^2r_2)}{D}$$

$$b_2=\frac{f_1f_2(r_1-r_2)^2+(f_1X_1-f_2X_2)(f_2X_1-f_1X_2)}{D}$$

where

$$D=f_1f_2\{(f_1^2r_1-f^2r_2)(r_1-r_2)+(f_1X_1-f_2X_2)^2\}$$

The required values are then given by

$$R_{13}=\frac{a_0a_1^2}{(a_0a_1b_1-a_0^2b_2-a_1^2)}$$

$$C_{14}=\frac{(a_0a_1b_1-a_0^2b_2-a_1^2)}{2\pi a_0^2a_1}$$

$$C_{12}=\frac{b_2}{2\pi a_1}$$

$$R_{11}=a_0$$

The impedance at any frequency may be found by

$$Z=\frac{a_0+ja_1f}{(1-b_2f^2)+jb_1f}=r+jX$$

The following is a procedure to determine the element values R_{11}, R_{13}, R_{15}, C_{12}, C_{14}, and C_{16} in Fig. 10.

Given are the impedances r_1+jX_1 at frequency f_1, r_2+jX_2 at frequency f_2, and r_3+jX_3 at frequency f_3.

The first step is to solve the following six equations (preferably by the step-by step elimination method[5]) for the values of a_0, a_1, a_2, b_1, b_2, and b_3:

$$a_0-a_2f_1^2+b_1f_1X_1+b_2f_1^2r_1-b_3f_1^3X_1=r_1$$

$$a_0-a_2f_2^2+b_1f_2X_2+b_2f_2^2r_2-b_3f_2^3X_2=r_2$$

$$a_0-a_2f_3^2+b_1f_3X_3+b_2f^2r_3-b_3f_3^3X_3=r_3$$

$$a_1f_1-b_1r_1f_1+b_2f_1^2X_1+b_3f_1^3r_1=X_1$$

$$a_1f_2-b_1r_2f_2+b_2f_2^2X_2+b_3f_2^3r_2=X_2$$

$$a_1f_3-b_1r_3f_3+b_2f_3^2X_3+b_3f_3^3r_3=X_3$$

Then find the values of G, H, M, and N from

$$G=\frac{a_1+\sqrt{a_1^2-4a_0a_2}}{2a_0}$$

$$H=\frac{a_1-\sqrt{a_1^2-4a_0a_2}}{2a_0}$$

$$M=a_0a_2b_1-a_0^2b_3-a_1a_2$$

$$N=a_0a_2b_1-a_0a_1b_3-a_2^2$$

The required values are then given by

$$R_{11}=a_0$$

$$C_{12}=\frac{b_3}{2\pi a_2}$$

$$C_{14}=\frac{GM-N}{2\pi a_0a_2\sqrt{a_1^2-4a_0a_2}}$$

$$C_{16}=\frac{N-HM}{2\pi a_0a_2\sqrt{a_1^2-4a_0a_2}}$$

$$R_{13}=\frac{G}{2\pi C_{14}}$$

$$R_{15}=\frac{H}{2\pi C_{16}}$$

The impedance at any frequency may be found by

$$Z=\frac{a_0+ja_1f-a_2f^2}{1-b_2f^2+j(b_1f-b_3f^3)}$$

References

1. Synthesis of Cable Simulation Networks, Robert W. DeMonte. *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 78, Nov. 1959, pp. 682–86.

2. Transmission Circuits for Telephonic Communication (book), K. S. Johnson. D. Van Nostrand Company, Inc., Princeton, N. J., 1925, chap. XI, pp. 121–42.

3. Impedance of Loaded Lines and Design of Simulating Compensating Networks, R. S. Hoyt. *Bell System Technical Journal*, New York, N. Y., vol. 3, July 1924, pp. 414–67.

4. Extensions to the Theory and Design of Electric Wave Filters, O. Z. Zobel. *Ibid.*, vol. X, April 1931, pp. 284–341.

5. Introductory Circuit Theory (book), E. A Guillemin. John Wiley & Sons, Inc., New York N. Y., 1953, chap. 3, pp. 112–20.

Santa Barbara County 6,000-Mc Microwave System

R. K. LENTZ
NONMEMBER AIEE

Summary: Important considerations in the design of a microwave system are power output from the transmitter, the antenna size, the distance from a terminal to the repeater, the path obstruction clearance, and the value of noise that will be permissible. The heart of this entire system lies in the method of interconnecting the radio equipment at the repeater site. The interconnection described in this paper was accomplished by back-to-back hybrid transformers rather than costly drop insert equipment.

EARLY in 1957, the General Telephone Company of California was faced with the problem of providing adequate telephone facilities for Vandenberg Air Force Base (AFB). This important defense and research complex is in an area of several thousand acres located about 10 miles east of the town of Lompoc in Santa Barbara County, California. It is bounded by the Pacific Ocean on one side and virtually uninhabited land on the other. During World War II, it was the site of Camp Cooke.

The increasing importance of Vandenberg AFB, particularly in connection with missile and other defense activities, indicated the desirability of providing ample communication facilities that could be quickly and economically expanded as needed. Initial studies disclosed that a communication network linking Vandenberg AFB with the existing telephone central offices in Lompoc, Santa Maria, Solvang, and Santa Barbara was essential for this purpose. The terrain lying between these locations is quite mountainous and frequently subject to forest fires. It was known, for example, that the local power company experienced considerable damage to their pole lines from forest fires in this area. In addition, the prevailing ocean winds would cause severe corrosion to open-wire lines that may be installed through this territory. Therefore, the best means of furnishing reliable communication facilities were considered

---

Paper 61-1113, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted April 14, 1961; made available for printing August 21, 1961.

R. K. Lentz is with the General Telephone Company of California, Santa Monica, Calif.

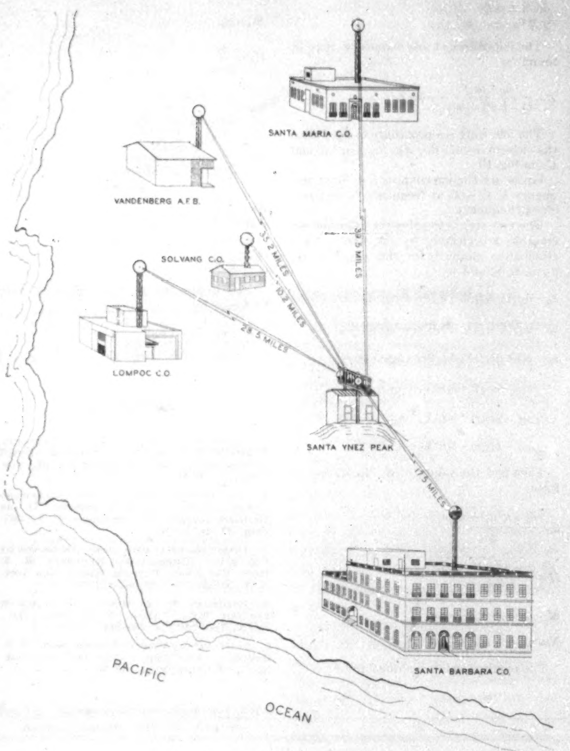


Fig. 1. Relationship of terminal points involved

to be underground cables and microwave radio systems.

Preliminary investigations revealed that underground cable would be the most economical approach when considering the circuit requirements between Vandenberg AFB and Santa Maria, a distance of approximately 20 miles. For the longer distance of about 70 miles to Santa Barbara, however, a microwave installation appeared to offer the most practical and economical solution. It was decided, after reviewing all pertinent factors, to provide a microwave system utilizing its inherent amplifying capability as the main means to interconnect Lompoc, Santa Maria, Solvang, and Santa Barbara with Vandenberg AFB. An existing aerial cable between Vandenberg AFB and Santa Maria would be used to satisfy alternate routing requirements.

Microwave Network

Map and terrain studies of the area pointed to the desirability of a central point, affording line-of-sight to all of the afore-mentioned towns, to serve as the site of a microwave repeater station. For this purpose, it was therefore deemed essential that the repeater point should possess each of the four following prerequisites:

1. Adequate height above the surrounding terrain to allow the use of short antenna towers for an unobstructed line-of-sight path to the terminal stations.
2. Ease of access by road to the site.
3. Availability of economical commercial power.
4. Economical cost for use of site as a microwave repeater station.

Selection and Evaluation of Repeater Site

Santa Ynez Peak, approximately 4,300 feet above mean sea level and located about 18 airline miles northwest of the City of Santa Barbara, was found to fulfill these requirements admirably. Fig. 1 shows the relationship between Santa Ynez Peak and the five terminal stations of the microwave network. There was a gravel access road leading to a fire lookout tower maintained at the top of Santa Ynez Peak by the U. S. Forest Service. Commercial power was readily available as power service was provided to a nearby television station. A special use permit was granted by the forestry service at a nominal cost; thus the four prerequisites for the repeater site were fulfilled.

Due to the need for a large number of voice-frequency channels, a broad radio-frequency bandwidth was dictated. The 6,000-mc (megacycle) band was chosen for the microwave system.

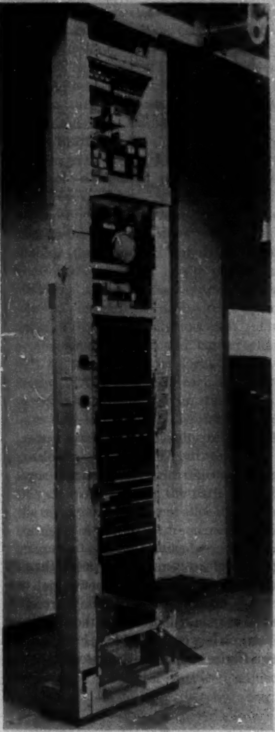
To evaluate the Santa Ynez Peak repeater site from a standpoint of radio propagation at 6,000 mc, profile maps to each of the five terminal locations were prepared. The free-space propagation loss, Fresnel zone clearance, and estimated antenna height and gain figures were computed for each path.

Service needs did not allow time to conduct actual 6,000-mc transmission tests to verify the preliminary map studies and computations. Furthermore, such tests would have had to be conducted over a sufficient period of time to determine propagation conditions during all times of the year, particularly the existence of seasonal fading effects. In this connection, a radio transmission survey of a week or two would be inconclusive and possibly misleading. The meteorological records of the area showed substantial diurnal and seasonal temperature variations. This indicated the likelihood of fading occurring in the various paths, particularly during the summer months. (Subsequent system tests revealed such seasonal fading effects. However, the fading effects that did show up were most severe at night between 1:00 a.m. and 8:00 a.m., during all seasons of the year.)

System Design

MICROWAVE SYSTEM PARAMETERS

The degree of propagation variations or fading predicted from the known meteorological conditions indicated the necessity of providing for frequency diversity operation on all the microwave circuits. Under normal operating condi-



tions, when using frequency diversity operation, the carrier input is multipled into two radio transmitters. These two transmitters operating on slightly different frequencies send out the same signal simultaneously. At the receiving end, the carrier output is derived from two radio receivers. Each receiver is tuned to one of the two transmitted frequencies and their in-phase combination yields a net improvement of 3 db (decibels) in the signal-to-noise ratio. For optimum service, it was deemed desirable that whenever the maximum permissible noise on a voice frequency channel approached 52 dba (db adjusted) with a signal input of 0 dbm (db referred to 1 millivolt) the frequency diversity transfer should be actuated leaving the quietest one of two receivers in operation. To insure adequately received field strengths over the longest paths with practical size parabolic antennas, it was determined that the microwave transmitter should have a power output of at least 1/2 watt in the 6.000-mc band. These criteria were fulfilled by Lenkurt type 74 microwave equipment and so this equipment was specified for the system.

Multiplex Equipment

After the type of microwave radio equipment had been determined, the next step in system design was to choose suitable multiplex equipment (carrier) that would meet the physical and electrical characteristics imposed upon it by the over-all system requirements. This included items such as maximum allowable per channel noise, channel bandwidth, channel frequency response, signaling requirements, physical size, and arrangements.

A design criterion was set whereby the maximum allowable average noise per channel on an end-to-end basis was to be below 29 dba for 50% of the time, and below 52 dba for 99.9% of the time.[1]

Fig. 2 (left). 6-wire patch bay

Test and Restoration Equipment

Having determined the design required to meet the microwave system objectives, it was then necessary to consider methods of maintaining and increasing the flexibility of the system. In order to minimize service interruptions, special equipment was required that would enable maintenance forces to patch circuits from one carrier channel to another without major interruption to important services. This special equipment would also allow monitoring of circuits on a bridging basis for test and measuring purposes and have built-in equipment capable of sending and receiving voice-frequency signals at levels required for routine maintenance. The particular equipment selected for this purpose was the 6-wire patch bay. It was designed with built-in features that would allow direct connection to special test equipment such as a pulsing test set, a noise-measuring test set, a variable frequency oscillator, db meters, and other carrier-type test equipment as required. Fig. 2 shows a 6-wire patch bay, Fig. 3 a simplified schematic of the 6-wire patch

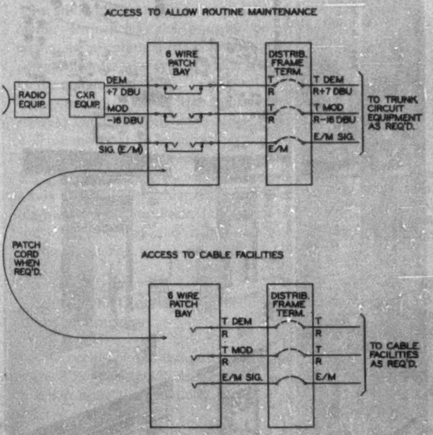


Fig. 3. Simplified schematic of 6-wire patch bay

Fig. 4. Santa Ynez repeater building

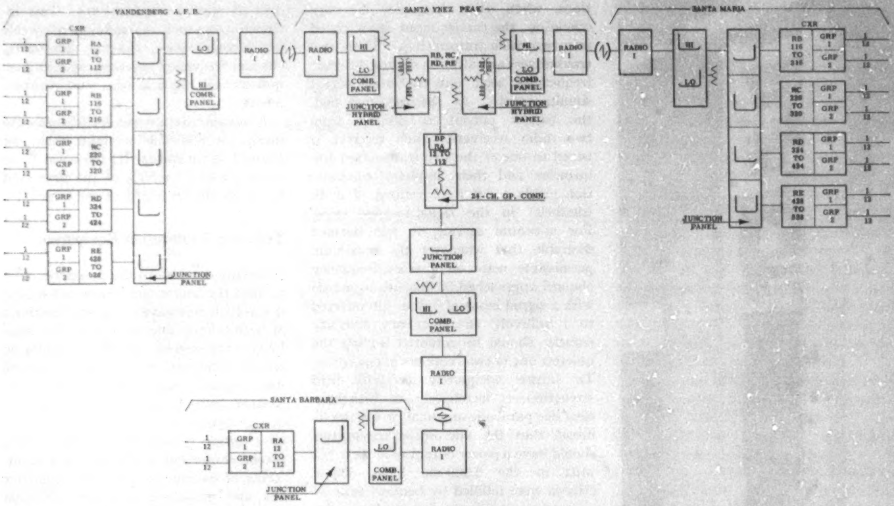

VANDENBERG A.F.B.
SANTA YNEZ PEAK
SANTA MARIA
SANTA BARBARA
CXR
GRP 1
GRP 2
1/12
RA 12 TO 112
RB 116 TO 216
RC 220 TO 320
RD 324 TO 424
RE 428 TO 528
JUNCTION PANEL
COMB. PANEL
HI
LO
RADIO I
RB, RC, RD, RE
JUNCTION HYBRID PANEL
BP RA 12 TO 112
24-CH. GP. CONN.

Fig. 5. System layout

bay utilizing two wires for the voice-frequency modulation leads, two wires for the demodulation leads, and two wires for pulsing and supervisory leads.

In addition to the 6-wire patch bay, a separate jackfield was installed. This jackfield provided access to the cable facilities on a 6-wire basis. This arrangement (see Fig. 3) allowed patching between microwave and cable derived circuits. With this equipment the system was flexible enough to meet varying military requirements and at the same time provide the means of restoring, maintaining, and rerouting circuits as required.

Santa Ynez Repeater Arrangement

The special use permit which was granted by the Forestry Service allowed use of a 40-foot by 80-foot plot of land 28 feet north of the fire tower at Santa Ynez Peak. The permit also authorized the construction of buildings and antenna structures. Fig. 4, Santa Ynez repeater building, shows the parabolic antennas, H-frame structure, lightning rods, and buildings. The antenna structure is located in such a position as to minimize the waveguide length between antenna and radio equipment. The small building in the background of Fig. 4 is used to house emergency power equipment which consists of batteries and charging equipment. A similar arrangement is used at the five terminal locations.

The antenna waveguides were connected to ferrite circulators which facilitated the connection of more than one transmitter-receiver unit to a common antenna. The ferrite circulator has an insertion loss of less than 0.5 db between ports and a reverse loss of more than 20 db. Ferrite circulators may be cascaded to connect transmitter-receiver units as required, providing the additional insertion loss is acceptable. For example, two ferrite circulators are cascaded at the

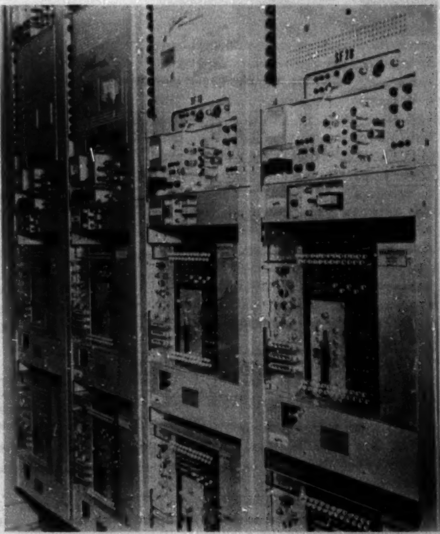


Fig. 6. Transmitter-receiver equipment at Santa Ynez Peak

Santa Ynez repeater site to allow the use of two radio systems, each having frequency diversity capabilities, in the path between Santa Ynez Peak and Vandenberg AFB. The same arrangement is used in the microwave paths to Santa Maria and Santa Barbara.

The heart of this entire microwave system lies in the method of interconnecting the radio equipment at the Santa Ynez Peak repeater site. All interconnections were accomplished by the use of wide-band back-to-back hybrid transformers which enabled the incoming signals from one location to be routed to the proper terminating office. This equipment was selected instead of a 4-wire bridging network because of its low cost. This arrangement also allowed carrier channels to be dropped for local service in groups of four channels each at the repeater site, should circumstances at a later date so dictate. The use of purely passive elements materially reduced the possibility of circuit outages. As a typical example of carrier routing through this complex consider the path taken by the repeater arrangement allocation which consists of 24 voice-frequency carrier channels between Vandenberg AFB and Santa Barbara as shown in Fig. 5. Fig. 5 depicts only a portion of equipment involved in this microwave complex.

In order to co-ordinate levels at the repeater site, attenuator pads were installed in both the transmit and receive legs. The attenuator pads were adjusted during system line-up, thereby assuring accurate levels for all carrier allocations when entering on the base band of the microwave radio terminals.

In addition to the interconnection of carrier frequencies, the repeater point serves as a common tie-point for the party-line order wire system which is used by the maintenance force.

Fig. 6 shows part of the transmitter-receiver equipment at Santa Ynez Peak which consists of 14 transmitters and 14 receivers serving five remote terminals. Each terminal has a transmitter and receiver arrangement comparable to that used at Santa Ynez Peak, but, in addition, the terminal also utilizes carrier equipment connected to the wire.

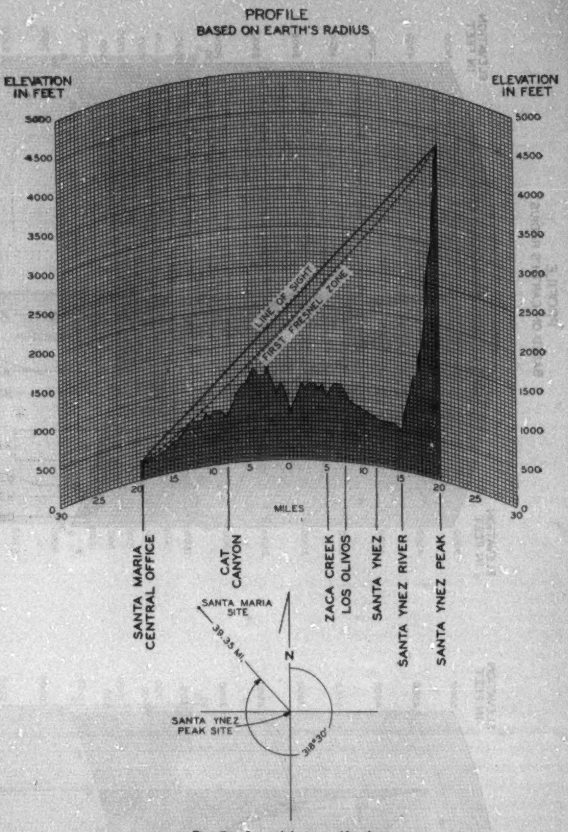


Fig. 7. Santa Maria profile chart

Profiles are based on earth's radius

Fresnel Zone Clearance Considerations

The results of the computation of the Fresnel Zone midpoint clearance for each location using the equation

$$F = 2{,}280\sqrt{\frac{d_1 d_2}{fD}}$$

where

f = frequency, mc
d_2 = distance from point in question to farthest terminal
d_1 = distance from point in question to nearest terminal
$D = d_1 + d_2$, total path length, miles

are as follows:

Santa Maria:	92.8 feet
Vandenberg AFB:	87.7 feet
Lompoc:	78.9 feet
Solvang:	46.8 feet
Santa Barbara:	58.3 feet

The midpoint Fresnel Zone clearance for each of the five microwave locations was adequate. A further examination of the profile maps revealed that for the Solvang path a possible near-end obstruction existed at the Santa Ynez Peak repeater site. This obstruction was a bare hill measuring 4,140 feet above average mean sea level and located approximately 0.45 mile from the peak. Although the Fresnel Zone clearance for the earth's radius, $K = 4/3$, was adequate, the clearance when using $K = 2/3$ of the earth's radius, is only 5 feet, which indicated the requirement for a tower approximately 200 feet high at the Solvang central office location. The 200-foot tower would represent approximately 5 feet of additional clearance over this obstruction. It was determined from an engineering economy standpoint that a 20- to 40-foot tower would be sufficient on the short path involved depending on

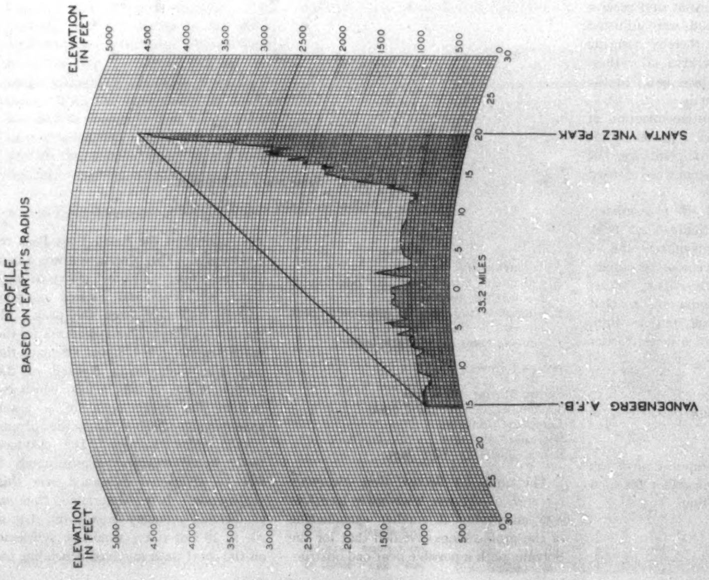


Fig 8. Vandenberg Air Force Base profile chart

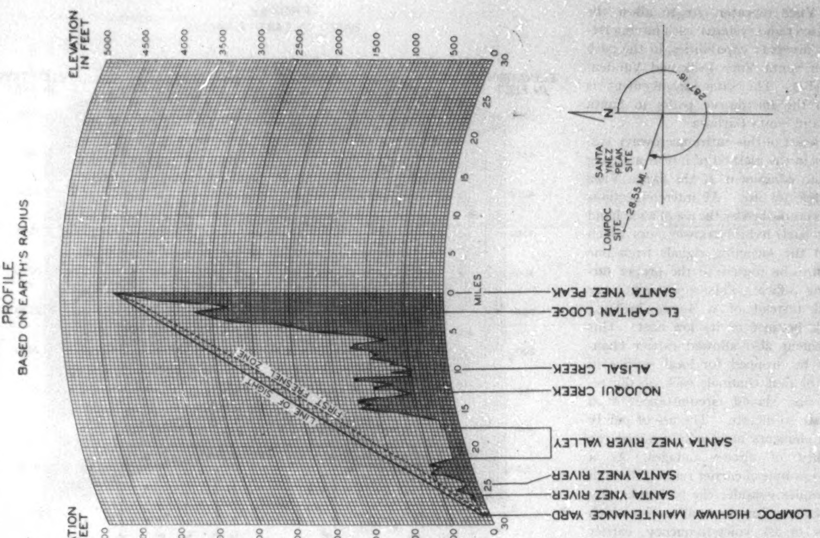


Fig. 9. Lompoc profile chart

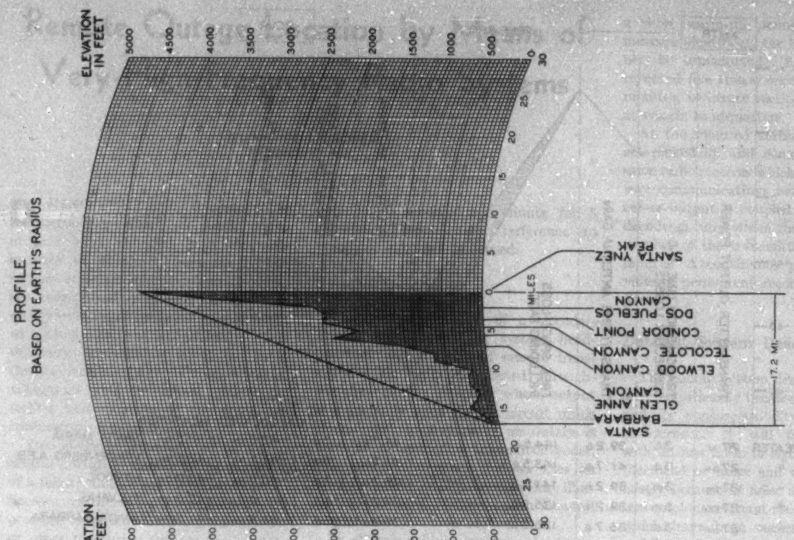


Fig. 11. Santa Barbara profile chart

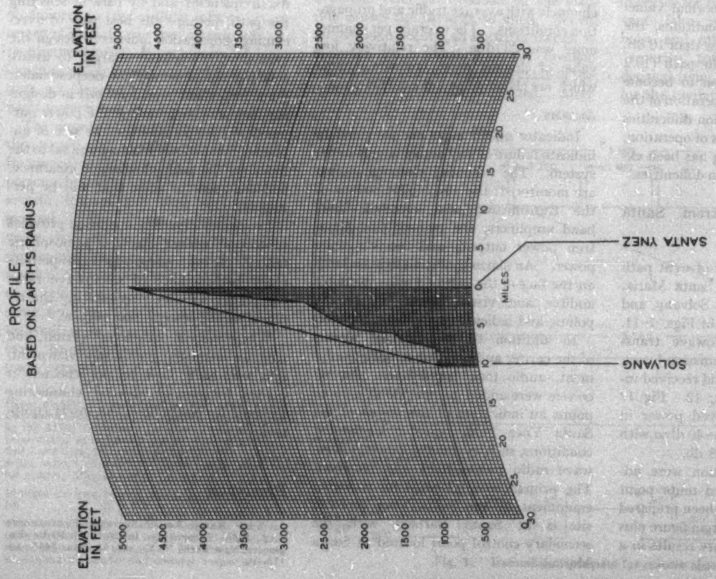
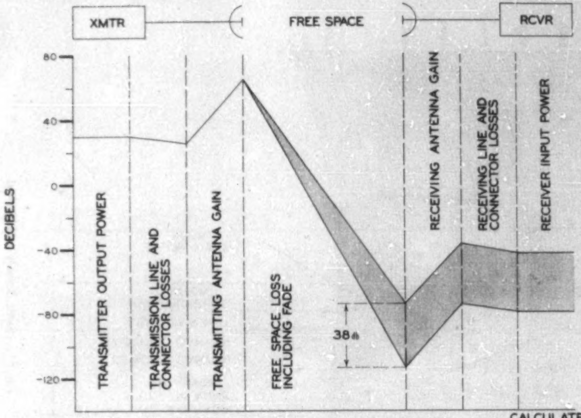


Fig. 10. Solvang profile chart



							CALCULATED	MEASURED	
SANTA YNEZ REPEATER	27 dbm	3 db	39.2 db	144.5 db	41.7 db	3 db	-42.6 dbm	-42.5 dbm	SANTA MARIA
" " "	27 dbm	3 db	41.7 db	143.5 db	41.7 db	5 db	-41.1 dbm	-42.5 dbm	VANDENBERG A.F.B.
" " "	27 dbm	3 db	39.2 db	141.0 db	39.2 db	3 db	-40.6 dbm	-40.0 dbm	LOMPOC
" " "	27 dbm	5 db	39.2 db	135.5 db	39.2 db	5 db	-41.0 dbm	-42.0 dbm	SOLVANG
" " "	27 dbm	3 db	35.7 db	136.5 db	39.2 db	3 db	-41.6 dbm	-41.5 dbm	SANTA BARBARA

Fig. 12. Radio system performance

the exact location of the antenna at the Solvang central office. It is true that the path would be below free space loss during normal propagation but limited to not more than 3 db below the normal value. Under two-thirds earth conditions, the obstruction loss would be less than 10 db. Considering the length of the path (10.2 miles) there was little reason to become concerned over the proper operation of the system because of propagation difficulties on this path. After 2 years of operation, no unusual degree of fading has been experienced due to propagation difficulties.

System Performance from Santa Ynez Repeater Site

Profile maps showing line-of-sight path from Santa Ynez Peak to Santa Maria, Vandenberg AFB, Lompoc, Solvang, and Santa Barabara are shown in Figs. 7–11. Values showing the microwave transmitter output power, transmission losses, antenna gain, space loss, and received input power are shown in Fig. 12. Fig. 12 also shows that the received power in each case is approximately −40 dbm with a fade margin exceeding 38 db.

Receivers at each location were adjusted to the recommended mute point of 52 dba. Statistics have been prepared that show that the fade margin figure plus diversity improvement figure results in a 99.9% reliability factor.[1]

End-to-End Performance

End-to-end average channel-noise measurements (*F1A* weighted) were made at the 4-wire drop side of the carrier derived channels with average traffic and propagation conditions. The average per channel noise was 23 dba. The relatively low value of channel noise provides circuits which are classified as toll-grade circuits.

Alarms

Indicator alarm units are provided to indicate failure of any active circuit in the system. The following electrical points are monitored: the 10-kc pilot output of the transmitting and received baseband amplifiers, the transmitting Klystron power output, and receiver noise power. An alarm condition lights a lamp on the face of the alarmed units, actuates audible and visual alarms at remote points, and indicates equipment transfer.

In addition to the internal alarms of the carrier and microwave radio equipment, audio tone transmitters and receivers were added to provide at remote points an indication of the status of the Santa Ynez Peak repeater equipment conditions, such as standby power, microwave radio transmitters, and receivers. The primary control point for remoting equipment at the Santa Ynez repeater site is at Santa Barbara with the secondary control point located at Santa Maria.

Conclusions

Because propagation variation cannot be controlled, except by choice of operating frequencies and by care in selecting the path, perhaps the best way of overcoming propagation fades is to design the system with adequate margin to assure a strong signal during the deepest fades. Some important considerations in designing such a system include the power output from the transmitter, the size of antennas, the distance from a terminal to the repeater, the path obstruction clearance, and the value of noise that will be permissible.

Frequency diversity operation provides protection against effects of atmospheric fading. It is a system which also permits ready maintenance of microwave radio equipment without necessitating the removal of the system from service.

Where system parameters permit, and from an engineering economy viewpoint, it is desirable to interconnect repeaters at carrier frequencies, thereby eliminating the need for costly drop and insert equipment.

Reference

1. UHF Radio Relay Systems Engineering J. J. Egli. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 41, Jan. 1953, pp. 115–24.

Remote Outage Location by Means of Very-High-Frequency Radio Systems

MARTIN COOPER
NONMEMBER AIEE

DURING THE PAST half century, the services provided by public power utilities have changed from desirable luxuries to absolute necessities. Electric power, water, and petroleum products are transported over long distances to satisfy widespread consumer requirements with a high degree of dependability. It is this very dependability which has made the occurrence of rare, but unavoidable, outages all the more serious. Rapid detection and location of the fault can minimize the loss and hazards.

Two methods are now used to alert a central operating station to the existence of a failure and its approximate location. Sensors, in the form of mechanically or pressure actuated switches, transducers, or other measuring devices are used to generate a signal which can be conveyed to the central station over electric cables. More commonly, because of the expense of running cables or renting telephone lines in large systems, the consumer has to be relied upon to telephone reports of failure to the utility office before remedial action can be attempted. A requirement exists for a system of fault location which is mechanized but which does not depend upon separate sets of wires to each sensor in a system.

The remote alarm system described in this paper uses existing vhf (very-high-frequency) radio networks to alert station personnel rapidly and automatically to system failures and their approximate location. The system operates over voice channels on 2-way radio circuits employing either carrier squelch or tone squelch. Radio transmitters are used at each remotely monitored station on the same frequency as the communications transmitter in the vhf network. However, transmissions occur infrequently, upon occurrence of a failure, and then only for 3 seconds per minute for 5 minutes. Thus, little interference to voice messages will be caused.

General Description

The system consists of a decoder-printer unit located at system headquarters, and a group of remote transmitter-coder units distributed throughout the system. A sensor whose output is a normally open electric circuit which closes upon occurrence of an outage is used to actuate the transmitter-coder. The description which follows uses an electric power utility system for illustrative purposes. Other applications will be discussed in a later section.

In electric utilities systems, the transmitter-coder units are located at unattended circuit breakers selected best to isolate a failure to a particular area. The transmitter is connected to the line side of a circuit breaker so that power is still available when an overload on the load side causes the breaker to open. When an overload occurs, an actuator switch is operated by a portion of the breaker mechanism, and the operating cycle of the transmitter-coder unit is started. After a short warm-up period, a tone-coded message identifying the breaker by number is transmitted. This message is repeated five times, once a minute for 5 minutes, to insure receipt of the message at system headquarters.

At the manned station the messages are picked up and demodulated by the same radio receiver which is used in the 2-way communications network. The receiver output is coupled into the digital decoder-printer which then types out the number of the transmitter reporting the failure. A continuous paper tape provides a permanent record of failure reports.

Detailed System Description

The coding system described here was selected primarily because of its reliability and its nonsusceptibility to noise and interference. It will operate reliably with an rms noise to signal ratio of 6 db (decibels) or more and will not operate falsely because of noise alone. The system employs resonant-reed relay detectors which have been successfully used on mobile and aircraft channels for many years.

Information on the location of a failure is coded in the form of two, three, or four serial decimal digits. Each digit is tone-coded using one of ten different tone frequencies for each decimal digit. To prevent false indications resulting from noise and other interferences, an 11th "security" tone is transmitted with each decimal tone. Fig. 1 shows the program for the carrier and tones associated with

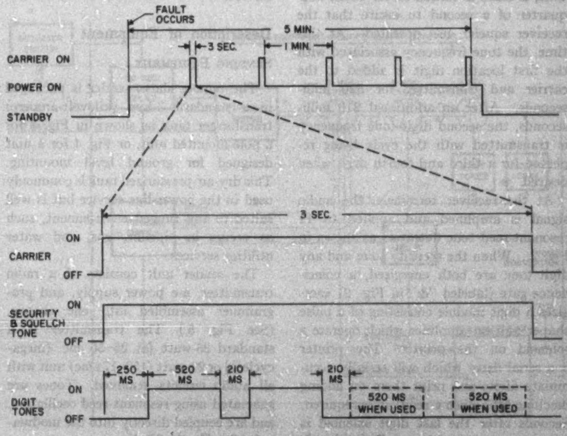


Fig. 1. Transmission program for remote alarm sender

Paper 61-920, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, Salt Lake City, Utah, August 23–25, 1961. Manuscript submitted June 13, 1960; made available for printing May 25, 1961.

MARTIN COOPER is with Motorola, Inc., Chicago, Ill.

The author wishes to acknowledge the excellent work of Frank Brugliera and Joseph Cardberg and their engineering groups at Motorola, Inc., in developing the sender and decoder, respectively.

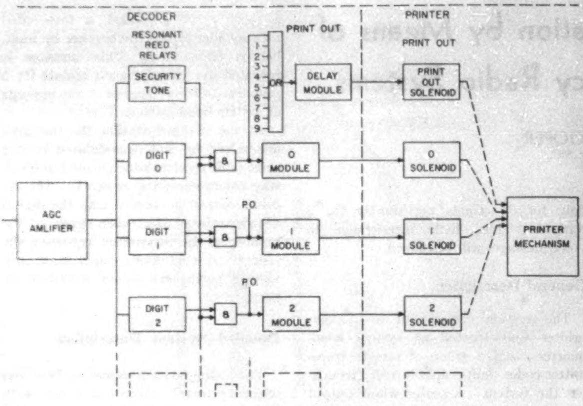


Fig. 2. Block diagram of digital decoder

a single failure. The squelch tone indicated in Fig. 1 is used only to make the remote alarm system compatible with systems using this form of squelch and is not required for the basic alarm system.

When a failure occurs, a pair of contacts is closed by the sensor at the point of failure and a programmer in the sender starts a program cycle. Once a program has started, it is not interrupted by any action of the operating contacts. After a warm-up period of about 55 seconds, the first of five identical 3-second transmissions starts. The carrier, modulated only by the security tone (and squelch tone, if used) is turned on for about one-quarter of a second to insure that the receiver squelch has operated. At this time, the tone frequency associated with the first location digit is added to the carrier and transmitter for 520 milliseconds. After an additional 210 milliseconds, the second digit tone frequency is transmitted with the cycle being repeated for a third and fourth digit when desired.

At the receiver terminals, the audio signal is amplified and applied to 11 resonant-reed tone detectors as shown in Fig. 2. When the security tone and any digit tone are both energized, a coincidence gate (labeled "&" in Fig. 2) energizes a digit module consisting of a pulse shaper and an amplifier which operate a solenoid on the printer. The printer is a serial lister which will serially accumulate digits and print them on adding machine tape when a read-out is required. Seconds after the last digit solenoid is operated, a delay unit energizes a read-out solenoid which causes the stored digits to be printed.

Thus, after each transmission is completed, a 2- to 4-digit number is recorded at the decoder. After five transmissions at 1-minute intervals, the sender stops and waits for the fault to be cleared. Correction of the fault will automatically reset the sender after a 2-minute period which insures that the correction is a permanent one. Succeeding failures will then cause the cycle to repeat. The transmission cycle described above is in accord with Federal Communications Commission regulations for this type of service.

Description of Equipment

SENDING EQUIPMENT

The remote alarm sender is packaged in a standard 3-kva (kilovolt-ampere) transformer tank as shown in Fig. 3 for a pole-mounted unit, or Fig. 4 for a unit designed for ground level mounting. This dry-air-pressurized tank is commonly used in the power-line service but is well suited to any rugged environment, such as occurs in pipeline, gas, and water utilities services.

The sender unit consists of a radio transmitter, a-c power supply, and programmer assembled into one chassis. (See Fig. 5.) The transmitter is a standard 25-watt [at 25–50 mc (megacycles)] or 20-watt (150–174 mc) unit with all audio circuits removed. Tones are generated using resonant-reed oscillators, and are coupled directly into the modulator at fixed preset levels.

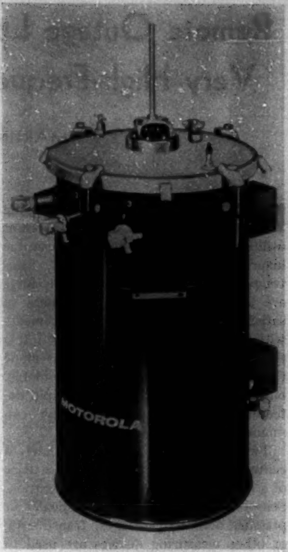


Fig. 3. Remote alarm sender, pole mounting

Programming is accomplished by a motor-driven gear train driving three cams at 20 rpm (one revolution per 3 seconds), 1 rpm, and 1/12 rpm (one revolution per 5 minutes). When the sender is actuated, snap switches on each cam perform the control functions necessary to complete a program cycle. The control functions include turn-on and warm-up of

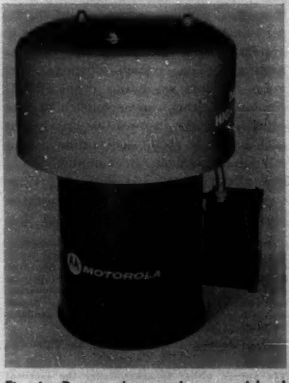


Fig. 4. Remote alarm sender, ground level mounting

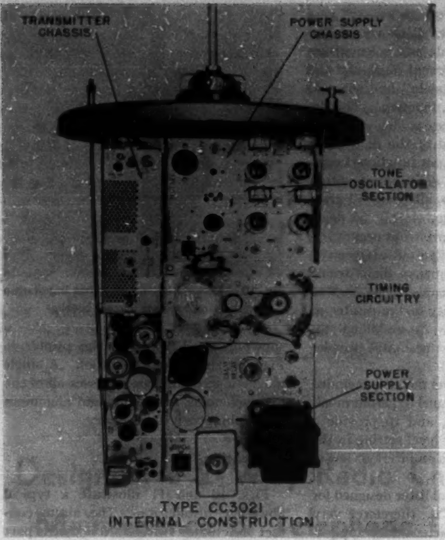

TRANSMITTER CHASSIS
POWER SUPPLY CHASSIS
TONE OSCILLATOR SECTION
TIMING CIRCUITRY
POWER SUPPLY SECTION
TYPE CC3021 INTERNAL CONSTRUCTION

Fig. 5. Remote alarm sender, interior view

the transmitter and resonant-reed oscillators, keying of the transmitter, and switching of the digit tones in accordance with the selected code. The oscillators operate continuously, and switching is accomplished by grounding all but the desired tones in the proper sequence. A block diagram of the sender is shown in Fig. 6.

The antenna and power connections are fed through the transformer tank via pressure-tight connections, and the vertically polarized antenna is supported at its connector in the center of the tank top, or separately mounted on the pole when the sender is located at ground level. In the case of pole-mounted units, four ground radials are directly connected to the tank top to form a ground plane for 25- to 50-mc operation, while the top itself acts as an adequate ground plane in the 150- to 174-mc band.

Selection of transmitter location is restricted to those areas where adequate signal strength can be obtained at the central station receiver. Limitations are thus imposed by both range and terrain. However, these limitations are negated in practical installations by the 20 to 30 feet of height gained by pole-mounting the unit. Since the radio network is designed primarily for mobile base communications, a high tower is usually used at the central station. Because of the higher antenna height provided by pole mounting and the smaller signal strength required for reliable signaling, coverage of the alarm sender system may be expected to be greater than that obtained from mobile radios in a system. Reliable operation was measured with signal strength of as low as 0.15 μv (microvolt) at a 160-mc receiver antenna input. The 20-db quieting sensitivity for the measured receiver was 0.35 μv. Field tests were conducted in an urban area using a 150-foot base station tower. The system was assumed to be operating "reliably" when 80 to 100% of the 3-digit location numbers were correctly received. Reliable operation was obtained over a 35-mile path where the rms noise voltage was 6 db greater than that of each of the two received tones. Since voice communication becomes marginal when signal is approximately equal to noise, voice signals were completely unreadable under this condition. Although a system would not normally be required to operate under these conditions, good practice dictates a reasonable reserve to account for variations in path attenuation and equipment gain.

The remote alarm power supply operates from a 117-volt 50- or 60-cycle-per-second power source, although direct operation of the coder and transmitter from a battery source is feasible.

Decoding Equipment

The digital recorder consists of a fully transistorized decoder driving a separate electromechanical printer. See Fig. 5. The signal input to the decoder is through a high-impedance bridging transformer so that no loading of the base station receiver will occur when the decoder is connected. A level control is provided to accommodate a wide range of voltages when the system is initially installed, but an automatic gain control (agc) amplifier permits signal input voltage variation of greater than ±10 db with no system degradation. The 11 resonant reeds, or Vibrasponders, are driven by the agc

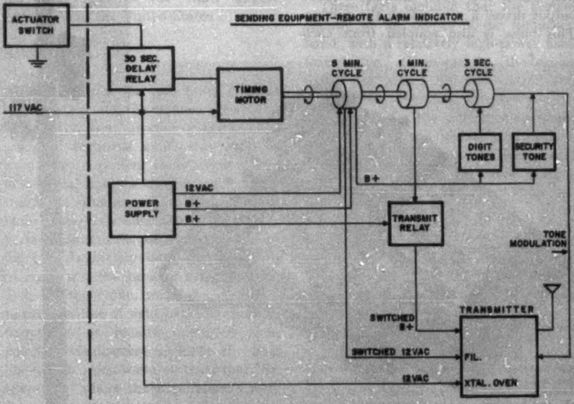

ACTUATOR SWITCH
SENDING EQUIPMENT—REMOTE ALARM INDICATOR
30 SEC. DELAY RELAY
117 VAC
TIMING MOTOR
5 MIN. CYCLE
1 MIN. CYCLE
3 SEC. CYCLE
DIGIT TONES
SECURITY TONE
B+
POWER SUPPLY
12VAC
B+
B+
TRANSMIT RELAY
TONE MODULATION
TRANSMITTER
SWITCHED B+
SWITCHED 12VAC
FIL.
12VAC
XTAL. OVEN

Fig. 6. Remote alarm sender, block diagram

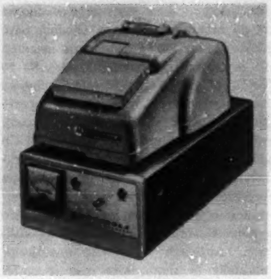


Fig. 7. Digital decoder and printer

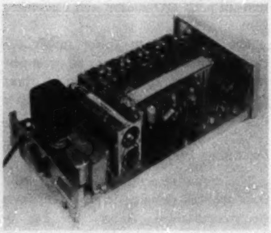


Fig. 8. Digital decoder, interior view

amplifier and contacts on the Vibrasponders provide the coincidence which operates the digit modules. Bandwidth of the reeds is sufficiently narrow to make the probability of false operation by noise very low. A properly received digit signal will cause two reed contacts to operate, producing a pulse which is reshaped and amplified by a 4-transistor trigger-amplifier, called the "digit module," whose output drives a solenoid on the printer. This pulse is also coupled from each digit Vibrasponder through an "or" gate into the "delay module" shown in Fig. 2. A separate module and solenoid are provided for each decimal digit, and the additional "delay module" and solenoid provide the print-out function. The delay module introduces a delay of $1^{1}/_{2}$ seconds from the end of the digit pulse after which the print-out function occurs. However, successive digit pulses within the $1^{1}/_{2}$-second delay interval will reset the print-out circuit and start a new delay period. As a result, print-out occurs $1^{1}/_{2}$ seconds after the last digit pulse is received regardless of the number of digits transmitted. Recording capacity is limited by the storage capacity of the printer to nine successive digits (or one billion stations). Fig. 7 shows the digital decoder and printer.

A transistorized level meter is included on the decoder front panel to permit monitoring of tone levels, and to provide a common reference for level setting in the senders. Fig. 8 shows an interior view of the digital decoder.

The printer is a serial lister designed for continuous use and is, therefore, very reliable in this intermittent duty application. The unique solenoid assembly shown at the left top of the printer in Fig. 9 is simple and foolproof. A single magnetic structure encompasses all of the digit solenoids, and individual alignment of the plungers is unnecessary.

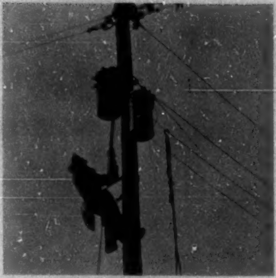


Fig. 10. Remote alarm powerline installation showing recloser circuit breaker

Applications

Figs. 10 and 11 illustrate a typical powerline installation. The alarm contact or actuator switch in this case is part of a recloser-type circuit breaker which

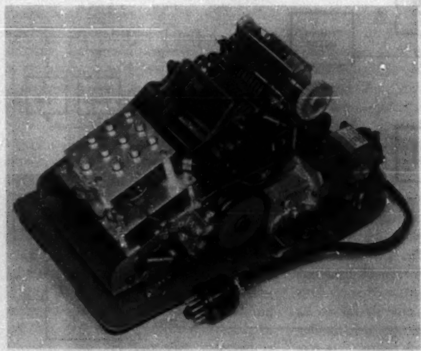


Fig. 9. Digital printer, interior view

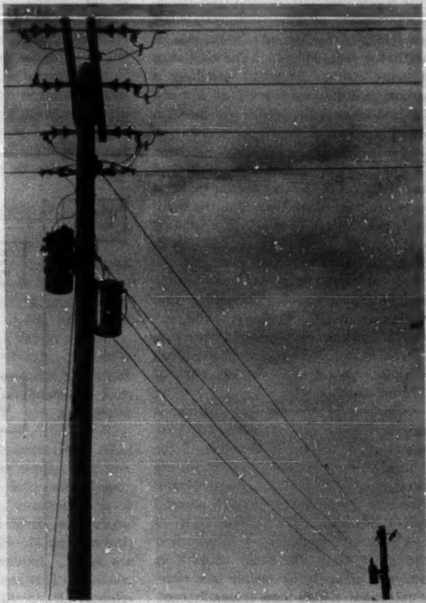


Fig. 11. Remote alarm sender installation

appears on the rear pole of Fig. 11. Operation of the circuit breaker closes the actuating switch which starts a transmit cycle in the remote alarm sender. A test switch is mounted on the nearby pole to permit manual initiation of a cycle for routine equipment check out. This installation also includes a lightning arrester for protection of the electronic equipment, and a 3-kva transformer to provide 117 volts a-c for the sender. Fig. 10 is a closer view of the sender installation showing the antenna and ground plane radials used with the low-band transmitter.

In the applications pictured, the units have fulfilled expectations. Within seconds of a real or simulated failure, an alarm has alerted station personnel and a permanent record of the failure location is in the hands of the repair personnel. Further use of the 2-way radio system then permits rapid dispatching of repair trucks to the location of the outage. Service is often restored before the consumer is aware of the failure.

Application to pipeline systems is essentially the same as that for electric power systems. It is necessary to have a usable communications path and a source of power at each sender location. Transducers may be used to operate on pressure, temperature, or other parameters which can be converted to a contact closure. In the absence of a utility pole, the antenna may be mounted on its own tower remotely located from the sender unit. Microwave or vhf relay stations will not interfere with system operation provided that no shifts in audio frequencies are encountered because of suppressed carrier single-sideband operation.

Conclusions

A system has been presented which combines efficient use of existing equipment and radio channels with simplified monitoring equipment to provide rapid reporting of outages in the power utility services. Interference with voice communications and cost of upkeep are minor. The permanent, easily read record of outage location will permit rapid remedial action to be taken with significant savings in repair manpower, and an improvement in customer convenience.

Design of a 2-Way Radio Channel for Maintenance of Telephone Company Services

D. L. MacDONALD
ASSOCIATE MEMBER AIEE

GADGETRY, all too often, has intrigued the application engineer. In his desire to try something new, he loses sight of his primary objective, which is to provide adequate communication in the most economical arrangement. It has been said that an engineer "is a man who can do with one dollar anything that any competent tinker could do with ten."

Purpose

In designing the system herein described, each idea was put to the test and made to prove itself on a cost-return basis. These cost studies are outlined, and the system and its usage are discussed.

Problem

How to design a radio system that would serve the needs of a large maintenance organization was a question asked by the Plant Department of the Pacific Telephone and Telegraph Company, located in Los Angeles. Many maintenance and construction functions are performed by this department, which also operates a privately owned motor-vehicle fleet, serving more than 600 square miles. The terrain varies from sea coast to mountains to desert, and the area includes the third largest metropolis in the United States.

Solution

When studies for this radio channel were being made, the engineer–designers, fortunately, were given free access to a group of technical administrators, who were able to present their requirements in a logical manner and, in turn, understand the advantages and limitations of radio as a communication medium.

As these studies progressed, the need for two completely independent groups of radio stations became evident. The first, a simple one, would be used car-to-car over line-of-sight paths to establish temporary or remote television pick-ups for broadcasters or make the final line-up on microwave stations and like operations where the requirements were for a quickly portable point-to-point network that might be in continuous service for long periods. For this 15-watt 450-mc (megacycle) station combination, an assortment of directional antennas were provided for use of television and microwave crews.

The second system, of more unorthodox design, is of principal interest. Its channel would be used initially by the construction forces. These crews are responsible for installation and maintenance—in isolated mountain canyons or in city streets—of all telephone company wire and cable between central offices and between central offices and individual subscribers.

The same equipment would form the backbone of our emergency communication network.

These facts dictated that the system operate in the 30- to 50-mc frequency band, with a relatively high-power fixed transmitter. Consequently, the choice fell to one with a 330-watt output and a 35.16-mc operating frequency. To compensate for the difference in power between fixed and mobile transmitters, multiple fixed station receivers were located throughout the area. Associated with this channel were three dis-

---

Paper 61-927, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Pacific General Meeting, Salt Lake City, Utah, August 23–25, 1961. Manuscript submitted May 25, 1961; made available for printing June 30, 1961.

D. L. MacDonald is with the Pacific Telephone and Telegraph Company, Los Angeles, Calif.

The described system is the result of close cooperation between the Secode Corporation and the Pacific Telephone and Telegraph Company. Mr. A. R. Ogilvie and Mr. W. R. Blake, representing the former, and Mr. J. J. Anderson and Mr. G. Applegarth, from the latter company, contributed significantly.

Fig. 1. Mobile unit

patch points, and the control terminal also could be used for dispatching in an emergency.

The antenna system associated with the transmitter is a locally designed colinear array, consisting of four half-wave elements driven in phase. It has a gain of about 5 db (decibels) relative to a dipole. This antenna also serves another 250-watt transmitter. The duplex arrangement has worked to our complete satisfaction.

The technical operator is stationed at the control terminal, which houses complete facilities for monitoring and operation of all base equipment. From this point, the operator can measure deviation and frequency of the base transmitter, select or reject permanently any receiver, and control audiolevels on the entire system. In addition, there are transmitter alarms to warn of low power, high-standing wave ratio, transmitter ON when it should be OFF, and transmitter OFF when it should be ON. There is also a light display, indicating normal channel operation.

To compensate for variations in audiolevel from each of the dispatchers, a voice-operated gain-adjusting device is included in the transmitter audiopath. This enables the system to be operated with a very high-average deviation and still exclude severe distortion due to excessive limiter action at the transmitter.

An automatic receiver selection circuit is used to accommodate multiple receivers. This circuit depends on a carrier-operated relay at each receiver site. An indication of relay closure is transmitted over a simplexed audiowire line to the control terminal. A relay circuit there selects the first responding receiver and rejects all others. A light display in the terminal allows the technical operator to determine which receivers responded to the signal and which was selected.

So that work operations of the dispatcher and mobile operator would not be interrupted, the system was made full-selective.

The most unusual items are found in the mobile units. Fig. 1 shows mobile control equipment, consisting of a standard station combination with squelch circuits disabled. Several vacuum tubes thus were eliminated, and a lowered power drain during stand-by operation resulted. There are no adverse effects since hang-up handsets are used rather than loud-speakers. In fact, the user gains an advantage in never being doubtful about the availability of the channel. If there is noise when the handset is off hook, the channel is idle; if no noise, the base transmitter is on and the channel is busy.

A system can be made full-selective in many ways, the well-known telephone dial being one illustration. This rugged simple instrument is perhaps the most sophisticated encoder ever developed. It has an address capacity that efficiently handles a nation-wide telephone system. Surely it will do for a simple dial-radio system.

When the handset is removed from its hanger, the mobile transmitter may be KEYED ON by moving the dial OFF NORMAL. As the dial returns to normal, a 2,805-cycle tone is turned on the number of times called for by the number dialed. This tone is transmitted from the mobile unit to equipment located at each dispatch point where the dial pulses are decoded. If the proper sequence of dialing tones is received, the decoder at the dispatch point signals the dispatcher.

Incorporated in the mobile control equipment is a device which allows the dispatcher to leave any one of five predetermined messages. This equipment is a circular cryptographic head with five jewels mounted in it. Which jewel is uncovered is determined by the particular code dialed by the dispatcher.

The importance of this feature was not fully realized until the work operation was analyzed. To be efficient, less than 20% of an operator's time should be spent in the vehicle. With this signaling arrangement, he is not recalled to his car to receive routine messages. A side benefit is a marked reduction in channel busy-time since repeated calls to a vehicle are never necessary—unlike with other signaling devices.

The decoder that makes all this possible is a special 8-function decoder. Each vehicle has a 3-digit address. This gives the system a theoretical capacity of 999 distinct addresses. However, one is used as a clearing pulse and is not available for an address digit. This and other restrictions reduce the number of addresses to about 700.

In operation, a dispatcher dials the mobile address; this causes a buzzer in the particular car to sound. If there is no answer, the dispatcher has two options. He may dial the digit, 4, which will either cause a loud bell to ring or, if at night, will light the car's spotlight. If the message is routine, he may dial the digit, 8, followed by either a 2, 4, 6, 8, or 0, depending upon which cryptographic light he wishes displayed in the mobile unit. The Pacific Telephone and Telegraph Company is not yet using this feature.

Further effort to reduce air-time makes extensive use of the TEN CODE. This has proved as useful in our service as it has in the police service, where it has long been standard. Using these methods, we are confident that at least 200 mobiles can be accommodated on the single-frequency pair.

At present, three dispatch points are

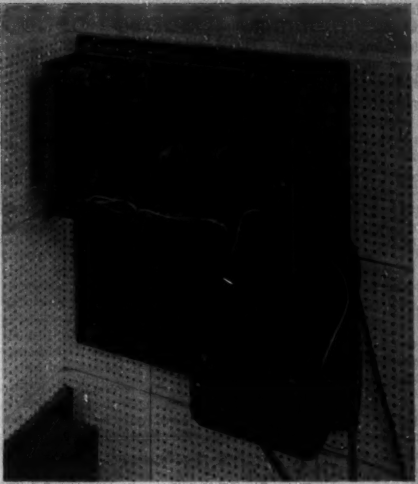


Fig. 2 (above). Land dispatch equipment

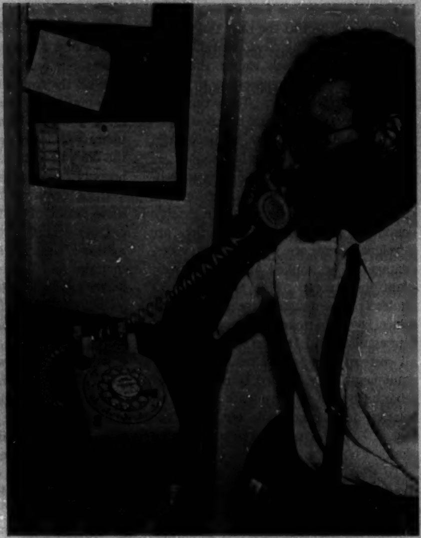


Fig. 3 (right). Dispatcher's desk equipment

used and all traffic is concentrated through them. Since the circuit is duplex and the mobile cannot turn on the base transmitter, mobile-to-mobile calls can be handled only through a dispatcher, thus placing him at all times completely in control of the circuit. This has resulted in a tight circuit discipline.

The dispatcher's equipment is housed in two small wall-mounted cabinets; see Fig. 2. They can be placed wherever convenient. Fig. 3 shows the handset, the only device on the dispatcher's desk. Within the wall cabinets are the power supply, station decoder, tone generators, and idle-line terminating circuits. The encoder is a high-impedance circuit and is across-the-line at all times. The decoder, also high-impedance, is across-the-line when the handset is ON HOOK and removed from the line in the OFF HOOK condition. This leaves the line unterminated in the ON HOOK condition unless an idle-line terminating circuit is provided. This is simply an impedance, which is placed across-the-line by the same relay that switches the decoder on and off the line. The value of this impedance is equal to the characteristic of the line. Thus, the bridging circuits at the control point always "look" into the same impedance, thereby, minimizing problems of level variation, singing, etc.

The transmitter-control and channel busy-lights are d-c circuits and operate over composite audiolines.

Economics

As previously mentioned, the engineer–designers of this system were forced to defend each item and each feature, making use of elementary traffic computation, work-procedure analysis, and cost studies. J. J. Anderson of the Plant Construction forces performed invaluable services, the cost and work-procedure studies being almost wholly his work.

Since the land-installation cost is distributed over all mobile units, the first step is to estimate the number of mobiles required immediately and 3 or 5 years hence. Such figures are not easily gained and, if the organization is new to 2-way radio, the estimates usually run spectacularly low.

A *detailed* analysis of work procedures is next prescribed, the emphasis being placed on the italicized word. In our particular situation, this study brought to light two separate requirements, which necessitated the development of two independent channels. The first—characterized by long intervals of inactivity, followed by periods during which the holding time is measured in hours; not minutes—would have raised havoc with any dispatch operation using the same frequencies. There are always many other questions to be resolved, the study generating the questions in each case.

With a firm estimate of how many mobiles are required, a look should be taken at circuit loading. A traffic engineer would be inclined to point out many items which should be considered and, in complex switched networks, these are extremely important. In our elementary system the computations illustrated in Appendix I are adequate. Holding time and calling rates usually must be estimated. Once the equipment is in operation, the data can be obtained with quite simple tools. The figures quoted are typical, but in nowise standard.

The number of mobiles, compared with the traffic estimate, will yield some knowledge of the grade of service. Channel loadings of 70% are acceptable; those exceeding 100% are common.

Because available frequencies are limited in number, the operator and engineer can do little to improve loading figures other than provide all possible means of reducing channel holding line.

Base station location, frequency, and power were selected in accordance with usual practices.

Costs quoted in the sample studies are

not to be taken as firm selling prices, but were in effect at the time this system was designed. The most desirable presentation is one which shows the daily or weekly time-saving necessary to justify the money expenditure. To do this effectively, several items must be known, i.e., total cost, number of units, annual charge, and loaded hourly rate for labor.

Total cost and number of units are self-explanatory. The figures for annual charges and loaded hourly rate deserve comment. Annual charges, as they are developed here, are presented as a percentage to be applied against any new plant investment. This percentage includes a proportionate share of all maintenance facilities, taxes, depreciation, and interest charges. In short, it should cover all costs of owning and maintaining the equipment. The loaded hourly rate for labor includes a proportionate share of all salaries, desk space, tools, etc. It is the cost of having a man perform an hour's work.

To use these figures, the cost of individual items is established. These are "installed" costs. The appropriate annual charge rate is then applied to the various items, and the dollar amounts thus obtained are summed. This amount, divided by the loaded hourly rate, gives the number of hours saving which must be effected in a year to make owning the equipment profitable.

In the case under discussion, a saving of 22.0 minutes per man per day would pay for the basic radio system; Appendix II details the computation. The next question to be resolved is the matter of full-selective operation. Here the computation shows that the dispatcher has to save only 4 minutes per day and the mobiles only 5 minutes per day. The price of the cryptographic device adds another 5 minutes to the required time savings at the mobile unit.

Totals show that the addition of radio can be justified by a saving of 33 minutes per day per mobile unit. The operators of this system estimate that 3 calls a week per mobile unit will save the required amount of time.

Conclusions

As an engineering study, this circuit presented a unique opportunity to investigate and propose new ideas. The channel has been in use through one rainy season and two major fires and has performed most satisfactorily, but further improvements are possible.

Use of a 2,805-cycle dialing tone from the mobile unit places an undue restriction on the connecting-wire line circuits. Automatic station identification by means of a code is being engineered. Since many vehicles remain stationary most of the day, all future mobile radio units will probably be transistorized.

In a LEAVE WORD signaling system, some doubt always remains in the dispatcher's mind as to whether or not the message was received. An interesting thought for further investigation is to make an unattended mobile RECEIPT a message. This is feasible because a decoder can become an encoder if supplied with a rate generator and a bit of control circuitry. Thus, any addressed mobile could RECEIPT a message by repeating address code back to the dispatcher and displaying the information on nixie tubes or some similar device.

A further interesting thought. Suppose that, in place of the simple *CRY-31* crypto unit, a 4-digit coding device is provided at both mobile and dispatch points. Ten thousand prearranged commands and replies are possible! These could be transmitted in an average time of about 12 seconds each, or 1,350 mobile units per channel. The channels would have to be spaced less than 5,000 cycles apart. Such a system probably could serve 8,100 units, where approximately 200 are served today. Perhaps voice communication is not the best way!

Appendix I. Traffic Considerations

Elementary methods are adequate for estimating traffic-handling capacities of simple systems. For the number of mobile units which can be accommodated, it is necessary to know:

1. Average channel time used per call. (Holding time.)
2. How many calls each mobile will generate during an average busy hour. (Calling rate.)
3. Desired percentage of time the channel is to be busy during each busy hour. (Per-cent load.)

$$\text{Number of mobiles} = \frac{60\times\text{Per-cent load}}{(\text{Holding time})(\text{Calling rate})}$$

In this instance,

Calling rate = 0.2
Holding time = 2 minutes
Per-cent load = 90%

$$\text{Number of mobiles} = \frac{60\times 0.9}{2\times 0.2} = 135$$

Appendix II. Cost Study

Annual charge = (Installed cost)(Annual charge rate).

Annual charges normally include:

1. Cost of money.
2. Depreciation.
3. Taxes.
4. Maintenance.

In brief, this charge includes all costs of owning a particular item, and is most easily used if converted to a percentage that can be applied to the installed cost.

In our case, the percentage charge to fixed and mobile radio equipment is 24% and 38%, respectively, as itemized, but variations will occur with different systems.

Fixed Radio Equipment	
$17,000 @ 24%	$ 4,080
Mobile Dispatch Type	
$65,000 @ 38%	24,700
Total annual charge	$28,780

Loaded hourly rate is $4.85

$$\frac{28,780}{4.85} = 5,934 \text{ hours which must be saved to pay for equipment}$$

$$\frac{\text{Hours of required saving}\times 3}{\text{Number of men}\times 13} = \text{number of minutes per day per man}$$

In our case

$$\frac{5,934\times 3}{63\times 13} = 21.7 \text{ minutes per man per day}$$

Similarily, there was additional charge for full-selective.

Installed cost of dispatch equipment was $400.

$$\$400\times 0.24 = \$96.00 \text{ per year}$$

$$\frac{\$96.00}{\$4.85} = 19.80 \text{ hours}$$

$$\frac{19.80\times 3}{13} = 4.6 \text{ minutes per day}$$

Installed cost of mobile units was $300.

$$\$300\times 0.38 = \$114 \text{ per year}$$

$$\frac{\$114.00}{\$4.85} = 23.51 \text{ hours}$$

$$\frac{23.5\times 3}{13} = 5.4 \text{ minutes per day}$$

Coded Feedback Communication Systems

J. J. METZNER
ASSOCIATE MEMBER AIEE

K. C. MORGAN
NONMEMBER AIEE

Summary: The conventional approach to coding for error correction is often impractical because of the required equipment complexity and the system vulnerability to changes in channel conditions, such as fading or burst-type noise. These problems can be solved by employing long codes with feedback, correcting only small numbers of errors, and utilizing appropriate feedback techniques. The resulting system yields essentially error-free performance under all conditions, is relatively efficient, and compares very favorably in cost.

AT THE PRESENT TIME a large gap exists between the state of development of coding theory and the utilization of this theory for digital data transmission. This is due in part, of course, to the relative newness of some of the theory, and to the time required for design and development. Two other factors, however, are more responsible for this gap, namely, the complexity of the equipment which would be required to implement all but the more elementary of these schemes, together with the inescapable fact that the reliability of any system protected solely by coding deteriorates seriously when confronted with fading or burst-type noise sufficiently strong to exceed the system thresholds. It is the assertion of this paper that recent developments in the theory of communication feedback with coding offer a potential solution to both of these difficulties at a cost which should prove to be acceptable in many situations.

Critique of Conventional Coding Capabilities

To provide perspective, a brief review of the pros and cons of previous coding theory is in order. Early results on coding indicated[1] that the price of an exponential reduction of error probability with code length at a nonzero information rate might be an exponential increase in computational complexity. Fortunately, the work of Elias[2] and Wozencraft[3] on iterative and sequential coding, respectively, has shown that the number of computations need go up only as a constant power of code length. Moreover, it is not essential to search for optimum codes. In sequential coding or in block coding a code may be chosen at random, followed only by a check to see that one of the few poor codes has not been chosen.

These coding techniques, however, are not completely satisfactory from a practical viewpoint if only unidirectional systems are considered. The following difficulties remain:

1. The coding-decoding schemes are far from simple when large error reductions are needed. Consequently, the cost of these systems tends to be formidable. An exception to this statement might appear to exist in the case of certain burst-correcting codes, which can be implemented in a fairly simple fashion. It must be realized, however, that these codes actually correct only a very limited set of error patterns, and consequently are of limited value.

2. Systems protected *only* by coding are inherently vulnerable to signal fading, burst-type noise (unless long bursts are prohibited), and, in general, sudden changes in channel conditions, since the transmitter cannot know when such changes occur. In such systems the error probability is determined roughly by the fraction of the messages for which the number of errors exceeds np_1, where p_1 is given by

$$R_t = 1 + p_1 \log_2 p_1 + (1 - p_1) \log_2 (1 - p_1) \qquad (1)$$

R_t being the transmitted information rate. When the errors are randomly spaced, Elias' bounds on error probability[4] apply. If the errors occur mostly in large groups, however, as with long bursts of noise or during fades, then the fraction of the time for which the number of errors exceeds np_1 is determined primarily by the frequency of the noise bursts or the fades. This frequency may easily be several orders of magnitude greater than the desired error probability. It can easily be shown that methods of avoiding this loss in reliability based on adjusting the system to operate satisfactorily under the *worst* channel conditions or to work against the average conditions are necessarily quite costly.

3. In addition to the uncertainty of the transmitter regarding the short-term variations of channel conditions, there is also uncertainty as to the average conditions over longer transmission periods. Since the reliability of any coded system deteriorates seriously when the system thresholds are exceeded, a wide margin of safety in information rate is usually required, even when some scheme such as interlacing is used to work against the average statistics.

4. In sequential coding, once an error has occurred all succeeding digits are received in error, which is inconvenient. Also, an unusually long computation may be required occasionally, which either increases equipment requirements or worsens error probability. Iterative coding requires a large amount of storage if the process is carried to more than two or three dimensions. Both systems, of course, require a sacrifice in information rate to achieve error protection.

A Solution: Coding Plus Feedback

The addition of a feedback channel, properly utilized, removes a good deal of the difficulty in each of the aforementioned categories. The system becomes more flexible, the decoding process becomes much simpler, and the capacity may be approached more closely. Furthermore, these advantages are realized while maintaining the reliability at a level unattainable, for practical purposes, with unidirectional systems.

These points can be brought out by considering an improved class of feedback systems which has recently been analyzed[5] in some detail by the authors. Essentially, the remarkable performance of systems of this type depends on three insufficiently appreciated facts:

1. When long codes are employed and error correction is attempted only for very small numbers of errors, excellent reliability is achieved. In fact, even if the signal is lost completely, as during a deep fade, the chance of accepting a code group which is incorrect can be made completely negligible.

2. A process of this type requires much less computation and storage than a corresponding system which attempts the maximum possible error correction.

3. Proper treatment of the feedback channel which reports accept-reject decisions back to the transmitter can make the effect of errors due to fading and disturbances in the feedback channel just as negligible as in the forward channel. This is accomplished by introducing a technique which permits detection of the fact that a code group has been repeated solely because an "accept" has been converted by noise into a "reject," and by utilizing a sufficiently asymmetrical decision mechanism at the transmitter to interpret the feedback information so that "reject" to "accept" errors occur with completely negligible frequency.

The resulting system yields excellent reliability, "fails safe" (an error probability of 10^{-10} when signal-to-noise ratio falls to zero, for instance), and operates at a fairly high per-symbol information rate. Because of its flexibility, the system is

---

Paper 61-1050, recommended by the AIEE Communication Theory Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted April 19, 1961; made available for printing August 15, 1961.

J. J. Metzner and K. C. Morgan are both with New York University, New York, N. Y.

The research reported in this paper was performed at New York University under the sponsorship of the Air Force Command and Control Development Division, Air Research and Development Command, Laurence G. Hanscom Field, Bedford, Mass., under contract no. *AF 19(604)-6168*.

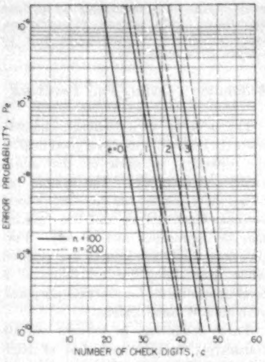


Fig. 1. Relation between error probability and number of check digits for several decision criteria and code lengths

particularly effective in precisely those situations where less efficient procedures fail, namely, when severe burst-type noise or heavy fading is encountered. The cost of this performance, in terms of computational and storage requirements, is much more reasonable than that of comparable unidirectional systems, and should prove to be acceptable in many situations.

The Forward Channel

Analysis of the forward channel falls quite naturally into three categories: the coding-decoding procedure, the resultant probabilities of error and rejection, and the requisite complexity of computation and storage.

THE CODING-DECODING PROCEDURE

An (n,k) group alphabet[6] (also called a systematic code or a parity check code[7]) is a set of 2^k n-digit code sequences in which k digits are specified as information digits and the other $c=n-k$ digits are check digits chosen to check parity over certain combinations of the k information digits. A group alphabet is defined uniquely by a matrix relating the check digits to the information digits. Thus, the c check digits $\{c_i\}$ are determined from the k information digits $\{i_j\}$ by

$$c_i=\sum_{j=1}^{k} m_{ij}i_j;\; m_{ij}=0 \text{ or } 1 \tag{2}$$

where the product $m_{ij}i_j$ is taken according to the ordinary rules of multiplication, while the sum is taken modulo two.

The first step in the decoding procedure for group alphabets is standard. It consists of the calculation of a parity sequence p_i determined by

$$\begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ \cdot \\ \cdot \\ p_c \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & \cdot\;\cdot & m_{1k} & 1 & 0 & \cdot\;\cdot\;\cdot & 0 \\ m_{21} & m_{22} & \cdot\;\cdot & m_{2k} & 0 & 1 & \cdot\;\cdot\;\cdot & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot\; 1 & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot\;\cdot & \cdot \\ m_{c1} & m_{c2} & \cdot\;\cdot & m_{ck} & 0 & 0 & \cdot\;\cdot\;\cdot & 1 \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ \cdot \\ i_k \\ c_1 \\ \cdot \\ c_c \end{bmatrix} \tag{3}$$

In the case where it is desired to make all possible error corrections the remainder of the decoding procedure is not well defined, but it appears that the amount of computation required for such a procedure is prohibitive for long codes.

When it is desired to correct only code groups containing very small numbers of errors, however, definite procedures which are reasonably simple can be specified. These are based on certain simple properties of group alphabets. When no errors are present in the received sequence, for example, the parity sequence consists entirely of zeros. When the received sequence contains only a single error in the i^{th} position, the resulting parity sequence is identical to the i^{th} column of the decoding matrix 3. It may be noted that if, and only if, the single error occurs in a check digit position the parity sequence contains only a single one. In general, if j errors are present in the received sequence, the parity sequence is the sum modulo two of the corresponding j columns of the decoding matrix.

For the type of system suggested in this paper it is desired to correct only code groups containing e or less errors, where e is a very small integer or, possibly, zero. Consequently, whenever the parity sequence calculated from a received code group consists of all zeros or is found to be identical to one of the

$$\sum_{i=1}^{e}\binom{n}{i}$$

modulo-two combinations of e or less columns of the matrix, the decoder performs whatever error correction is indicated and records the information contained in the code group. If more than e errors are found to have occurred, it rejects the code group without attempting further error correction.

THE PROBABILITIES OF ERROR AND REJECTION

The exact error probability depends on the particular code chosen. Determination of an optimum code for code lengths of 100 or 200 digits is effectively impossible with present techniques. Fortunately, it is not necessary to do this, since any parity check code chosen at random is almost certain to have behavior which is essentially as good as optimum.

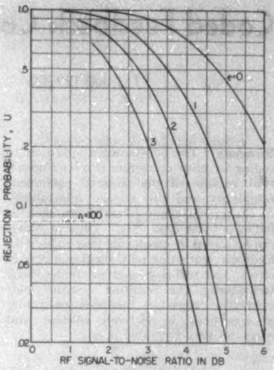


Fig. 2. Relation between rejection probability and radio-frequency signal-to-noise ratio for several decision criteria

n=100

Consider first what would appear to be the worst situation, received sequences that are completely random, bearing no relation to what was sent. For e-error correction the probability P_{er} that a random received sequence be erroneously accepted is then readily seen to be

$$P_{er}=\frac{1}{2^n}\sum_{i=0}^{e}\binom{n}{i}(2^k-1)$$
$$\doteq 2^{-c}\sum_{i=0}^{e}\binom{n}{i} \tag{4}$$

For small values of e, P_{er} is extremely small. When $e=1$, $n=100$, and $c=40$, for example, the error probability is 10^{-10}. For most practical purposes, an error probability of this magnitude is completely negligible. More complete analysis[5] shows equation 4 to be an upper bound on the error probability for most randomly-chosen codes.

A plot of equation 4 is shown in Fig. 1. It is seen that even with moderate numbers of check digits extremely low error probabilities are attainable. For the ranges of n and e shown on Fig. 1, for instance, an error probability not exceeding 10^{-10} requires only from 16 to 50% of the digits to be check digits. These various cases correspond to transmitted information rates of from 0.5 to 0.84, so that the coding procedure is seen to be fairly efficient. Of course, the system rate is somewhat lower than the transmitted

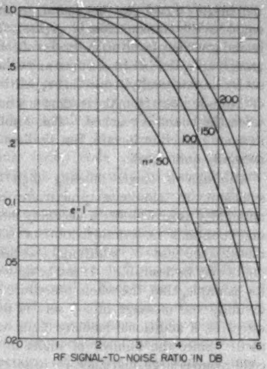


Fig. 3. Relation between rejection probability and radio-frequency signal-to-noise ratio for several code lengths

Single-error correction

rate because some of the code groups may be rejected one or more times, depending on channel conditions. It may be noted that for a given error probability the required number of check digits varies only slightly with n, as would be predicted from equation 4. Also, every addition of ten check digits yields about three orders of magnitude improvement in the error probability.

The probability that a received sequence will be rejected is, to a close approximation, just the probability that the sequence contains more than e digit errors. When digit errors occur at random with probability p_o, the rejection probability U is, simply,

$$U=\sum_{i=e+1}^{n}\binom{n}{i}p_o^{\,i}(1-p_o)^{n-i}$$

$$=1-\sum_{i=0}^{e}\binom{n}{i}p_o^{\,i}(1-p_o)^{n-i} \qquad (5)$$

Figs. 2 and 3 show the variation of U with signal-to-noise ratio for various combinations of n and e. For convenience, a synchronous-detection model has been assumed for which

$$p_o=\frac{1}{2}\left[1-\frac{2}{\sqrt{\pi}}\int_0^A e^{-\alpha^2}d\alpha\right] \qquad (6)$$

where A^2 is the radio-frequency signal-to-noise ratio. The curves exhibit a definite threshold effect. The transition from fairly high to fairly low rejection probability occurs over a range of about 2:1 in signal-to-noise ratio. Thus, when fading occurs, the rejection probability is roughly equal to the fraction of the time during which the signal-to-noise ratio is below a certain value. This, however, is far better than the situation for any practicable nonfeedback system, for which the error probability would be determined roughly by the fraction of the time during which the signal-to-noise ratio falls below a certain value. The critical signal-to-noise ratio is somewhat lower in the latter case, but the difference is slight.

The effect of the rejection threshold can be seen in Fig. 4. Here, Rayleigh fading is assumed, and the average rejection probability at the receiver is plotted as a function of the average signal-to-noise ratio for $n=100$ and single-error correction. The $e=1$ curve of Fig. 2 indicates that the rejection probability is very small for signal-to-noise ratios appreciably greater than 5 decibels (db) and approaches unity rather abruptly below this value. This is reflected in the fact that Fig. 4 shows an average rejection probability of about 0.5 when the average signal-to-noise ratio is 5 db.

In an actual communication channel digit errors do not usually occur entirely at random. Even if there is no fading, noise tends to occur in bursts, so that the probability of a large number of errors in a code group is greater than that predicted if a binomial distribution is assumed. In an exclusively error-correcting procedure the frequency and duration of such bursts set an upper limit on the attainable reliability unless some sequential type of coding is used for burst-error correction. In the partial error-correction procedure described here (which should prove much less costly) reliability is not affected, and, in fact, the fraction of rejected code groups is usually less if the noise occurs in bursts than if it is uniform.

The Complexity of Computation and Storage

In this section the number of binary comparisons per bit will be taken as a measure of the computational complexity. By definition, the modulo-two sum of two binary digits is a single binary comparison. The modulo-two sum of j binary digits will be considered to be equivalent to j binary comparisons. Obviously, this approach by no means yields an exact measure of the computational complexity. It is, however, sufficiently meaningful to allow order-of-magnitude comparisons.

In coding, each check digit is the modulo-two sum of about half of the information digits, and thus requires about $k/2$ binary comparisons for its calculation. This corresponds to about $c/2$ binary comparisons per bit of information. In decoding, each parity digit is obtained by the modulo-two sum of one check digit and about half of the information digits, so that about $c/2$ binary comparisons per bit of information are needed to determine the parity sequence.

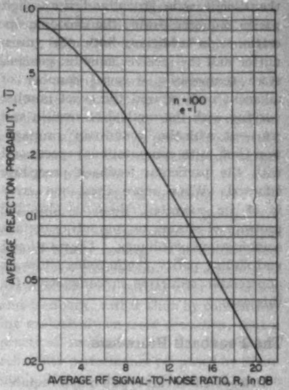


Fig. 4. Relation between average rejection probability and radio-frequency signal-to-noise ratio with Rayleigh fading

Single-error correction
$n=100$

For zero-error correction no further computation is required. It can be shown,[5] however, using arguments based on the decoding procedure outlined directly after matrix 3, that single-error correction requires only about two additional binary comparisons per bit. Double-error correction can be shown[5] to require about k additional binary comparisons per bit. Also, in this case, $k(k-1)/2$ pairs of columns must be either stored or computed. The number of computations required to correct greater numbers of errors increases rapidly with the number of errors if the procedure described previously is extended directly. The authors are not aware of any simpler general procedure, although particular codes may offer possibilities of simplification.

Associated with the decoding computations are certain storage requirements. For general parity check codes a coding matrix of kc digits must be stored permanently at both the transmitter and the receiver. However, by using a sliding parity check code,[4] the permanent storage requirement is reduced to $n-1$ digits, and if a rotational code[5] is used, it is reduced to only k or c digits, whichever is greater. The receiver must also store the n-digit code group and the c-digit parity sequence during the decoding operation. In addition, both the transmitter and the receiver must, in general, store temporarily a small number of adjacent code groups, the exact number varying with the method of read-in and read-out, with the "round-trip" propagation time, with computational speed, and with the particular feedback procedure adopted. When more than one error is to be corrected, there is also the problem of storing or computing certain combinations of columns. The rotational and sliding parity check codes allow an appreciable reduction in complexity for this case.

The Feedback Procedure

In a system of this type a major requirement is the selection of a reasonably simple feedback procedure which prevents any possible type of error from having an appreciable effect on reliability. In this section such a procedure is outlined for both unidirectional and bidirectional message flow.

For every code group in the forward channel, one bit of information is sent from receiver to transmitter via the feedback channel to indicate whether the message has been accepted or rejected. Two types of feedback errors can occur: an acceptance can be interpreted as a rejection ($A \rightarrow R$), or vice versa ($R \rightarrow A$). It has been pointed out in previous studies that, because of the lower information rate required, the feedback channel can often be made considerably more reliable than the forward channel. When fading or severe noise bursts are present in the feedback channel, however, it is unlikely that the probabilities of both types of error can simultaneously be reduced much below the limit set by the frequency of such interruptions. Fortunately, the effect of each type of feedback error can be rendered negligible by employing a different stratagem for each.

An ($R \rightarrow A$) error may result in the loss of one or two code groups. The probability of this happening can be made completely negligible if the transmitter uses a sufficiently asymmetrical decision mechanism to interpret the feedback signal. Thus, as one example, the receiver may send one of two sequences of m digits to represent an acceptance, and another to represent a rejection. The transmitter should then read the feedback as an "accept" only when the sequence it receives is within d digits of the appropriate sequence, where $d << m$. If $m = 40$ or 50, the ($R \rightarrow A$) error probability will be less than 10^{-10} even during complete fades.

A necessary consequence of the previously outlined procedure for eliminating the problem of ($R \rightarrow A$) errors is that ($A \rightarrow R$) errors will occur with appreciable frequency whenever channel conditions are poor. Hence, the receiver needs some means of ascertaining whether an incoming code group actually contains a new portion of the message or is merely an undesired repetition of a previously transmitted and accepted group. Techniques for providing this information are readily available.

The easiest situation to handle is that in which a new code group is not sent out until the corresponding feedback message is received. In this case it is only necessary to reserve one digit position in each forward channel code group, filling it alternately with a 0 or a 1 in successive new code groups. When a repeated code group which has not been requested is accepted by the receiver, the special "labeling" digit is seen to be the same as that in the preceding recorded code group. The unrequested repeat is thus recognized, and may be "crossed out."

The problem is somewhat more complex if forward and feedback signals are sent simultaneously on separate channels. Here it is necessary to send some number, $j-1$, of additional code groups before it is known whether the first has been accepted or must be treated as though rejected. The value of j depends on the round trip propagation delay time, the decoding time, and the duration of the feedback signal. For most actual system configurations these delays are small compared with the length of a code group, and $j = 2$.

One way of solving this problem requires the use of only one labeling digit. Assume that the entire message sequence in binary form is encoded into groups in preparation for transmission. These groups are divided into j sets, say $A_1, A_2, \ldots, A_j$, in the following manner. The set A_1 contains the first, $(j+1)^{\text{th}}$, $(2j+1)^{\text{th}}$, ... groups and, in general, A_i contains the i^{th}, $(j+i)^{\text{th}}$, $(2j+i)^{\text{th}}$, ... groups. The first transmission is, then, the first code group from A_1, the second the first from A_2, and so on to A_j. The $(j+1)^{\text{th}}$ transmission is either the second code group from A_1 or a repeat of the first, depending on the feedback message relating to the first transmission. Thus, the transmission of each of the j sets proceeds, in effect, independently of the others, although interlaced as described. A storage problem arises in some situations, because the cumulative numbers of requests for repetition in the j sets do not, in general, remain equal as the transmission progresses. A bound must be placed on the excess of code groups accepted in one set over any other set by automatically requesting repetition of the code group from the leading position when the bound is reached. This results in some reduction in rate, but the loss is normally quite small.

Although a single labeling digit is sufficient to eliminate practically all ambiguity as to whether a decoding message contains new information or is a repeat, the use of additional labeling digits may be desirable. It can be shown, for example, that the whole procedure of dividing the message into j sets is not essential, if additional labeling digits are used. Since the number of digits in a code group is large, and since, for instance, 5 digits are sufficient to distinguish 32 successive code groups, this additional labeling can be done with little loss in rate.

The use of additional labeling digits also provides improvements in reliability under certain conditions. For instance, when only one labeling digit is used, a single ($R \rightarrow A$) error, although extremely rare, can cause the omission of two successive code groups without the receiver being aware of the omission. Such an error can be serious in some communication systems. By employing two labeling digits instead of one, however, only one code group need be lost, and its position is known by the receiver. Even this loss can be avoided, but the cost of the additional logic circuitry involved is probably not justified by the slight gain. As another example, when using the procedure described above in which j sets of code groups are distinguished, a prolonged loss of synchronization might cause confusion as to which set A_i is currently being received immediately upon restoration of synchronization. Additional labeling digits can readily be shown to avoid this possibility of confusion. It is even possible to label every code group in a long message sequentially, if the somewhat larger loss in rate can be tolerated. Such an extreme is not likely to prove desirable, however. The use of additional labeling digits also suggests the possibility of other feedback procedures, but space does not permit discussion of them here. A number of methods of employing feedback have been described in previous analyses.[8-14]

The discussion of feedback to this point has dealt exclusively with the case where accept-reject information is returned to the transmitter via a signal constructed solely for this purpose. When message information is flowing in only one direction, some such arrangement is, obviously, a necessity. If information is to be transmitted in both directions, however, then feedback can be used in both directions, and the proportion of the channel taken up by accept-reject information can be made so small as to be almost negligible. Thus, if A and B represent the two stations, (n_1,k_1) code groups with e_1-error correction can be used to transmit from A to B, and (n_2,k_2) code groups with e_2-error correction from B to A. Only one digit in each code group is needed to carry the accept-reject information regarding the appropriate code group received in the other channel. Each time a message is rejected by A, A sends a repeat of its own previous message, along with a request for repetition of the previous B-to-A message. If, on the other hand, a message is accepted by A, there is almost no chance of error, so that A may send either a new message or a repeat, depending on the received accept-reject digit. There may, of course, be situations in which it is desirable to keep all the accept-reject information separate from the message-bearing code groups, even though the information flow is bidirectional.

By utilizing these techniques, the system reliability is rendered practically impervious to feedback channel disturbances. When conditions in either or both channels fall below the effective system thresholds, the only effect on performance is to reduce the information rate temporarily. Reliability is unaffected. Hence, to obtain practically error free transmission at a fairly high per-symbol information rate, it is only necessary to select the system parameters so that the probability of a code group containing more than e errors as received is not excessive.

An Illustration of System Performance

An example has been worked out[5] that illustrates the performance of a system based on the ideas described herein quite effectively. Consider a long-code binary communication system employing decision feedback with single-error correction. Assume that information is transmitted in both directions via (100,60) code groups. Also assume that the following channel conditions exist:

The channel statistics are the same in both directions.

Fifteen % of the code groups are transmitted when channel conditions are very poor, and thus contain large numbers of errors, while the other 85% of the code groups are transmitted at a 6-db signal-to-noise ratio.

The system performance can be described in terms of the error probability and the overall information rate. The error probability for the average code is better than 10^{-10} under these conditions. For a data rate of 1,000 bits per second (about 17 code groups per second) this corresponds to an error about once every 16 years. Equally good or better reliability could also be achieved under other channel conditions.

Assuming that rejections occur at random during the 6-db signal-to-noise ratio periods and mostly in long strings when conditions are poor, the ratio of the number of transmitted code groups to the number of accepted code groups is found to be 1.22. This means that, for every 100 transmitted digits, $(1/1{,}22)(58) = 47.5$ information digits are recorded at the receiver.

It should be noted that a nonfeedback, interlace coding system, if used on this channel, would have to work against an average error probability

$$P_{\text{avg}} = (0.15)(0.5) + (0.85)(0.0023) \doteq 0.077$$

This corresponds to a channel capacity of 0.61 bit per symbol. Such an interlace system would be prohibitively complex, so that it is extremely unlikely that such an arrangement could be used as a means of approaching the channel capacity. On the other hand, the rate of 0.475 bits per symbol for the decision-feedback system, which is practically error-free, is quite close to the maximum information rate for systems which operate on the average statistics of the channel. The comparison does, however, neglect the time which must be allotted in the feedback system for propagation delay and computation. This may or may not be of significance, depending on such factors as transmitted bit rate, length of the transmission path, and computing speed. If transmission is carried out simultaneously in both directions, this loss in rate can be avoided.

Conclusions

The type of coded decision-feedback system described herein yields almost perfect reliability at a rate which is an appreciable fraction of channel capacity regardless of the magnitude or stability of the channel signal-to-noise ratio. No practicable system without feedback can yield comparable performance. The feedback permits the information rate to adjust quite efficiently to the varying channel conditions, while the use of long codes with only partial error correction provides excellent protection against error at reasonable cost. New feedback procedures eliminate practically all errors arising from various types of noise in the feedback channel.

The coding-decoding complexity for single-error correction is shown to be particularly simple, so that this decision criterion appears to be preferable in many situations to the correction of larger numbers of errors.

The advantages of utilizing feedback in conjunction with long code groups have not been fully appreciated by investigators of binary coding. The authors feel that it is by emphasizing error *detection*, in conjunction with partial error correction and feedback, rather than by concentrating exclusively on error *correction*, that a practical realization of the very high-reliability performance promised by Shannon's Noisy Channel Coding Theorem will be achieved.

References

1. The Mathematical Theory of Communication, **C. E. Shannon.** *Bell System Technical Journal*, New York, N. Y., vol. 27, 1948, pp. 379–423, 623–56.

2. Error-Free Coding, **P. Elias.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-4, Sept. 1954, pp. 29–37.

3. Sequential Decoding for Reliable Communication, **J. M. Wozencraft.** *National Convention Record*, Institute of Radio Engineers, vol. 5, pt. 2, 1957, pp. 11–25.

4. Coding for Noisy Channels, **P. Elias.** *Ibid.*, vol. 3, pt. 4, 1955, pp. 37–46.

5. Reliable Fail-Safe Binary Communication, **J. J. Metzner, K. C. Morgan.** *WESCON Convention Record*, Institute of Radio Engineers, vol. 4, pt. 5, 1960, pp. 192–206.

6. A Class of Binary Signalling Alphabets, **D. Slepian.** *Bell System Technical Journal*, vol. 35, 1956, pp. 203–34.

7. Error Detecting and Error Correcting Codes, **R. W. Hamming.** *Ibid.*, vol. 29, 1950, pp. 147–60.

8. Theory of Information-Feedback Systems, **S. S. L. Chang.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, vol. IT-2, no. 3, Sept. 1956, pp. 29–40.

9. Binary Symmetric Decision Feedback Systems, **Bernard Harris, K. C. Morgan.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, Sept. 1958, pp. 436–43.

10. Optimum Decision-Feedback Systems, **B. Harris, A. Hauptschein, L. S. Schwartz.** *National Convention Record*, Institute of Radio Engineers, pt. 2, 1957, pp. 3–10.

11. Binary Decision-Feedback Systems for Maintaining Reliability under Conditions of Varying Signal Strength, **B. Harris, A. Hauptschein, K. C. Morgan, L. S. Schwartz.** *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 13, 1957, pp. 126–40.

12. Cumulative Binary Decision-Feedback Systems, **S. S. L. Chang, B. Harris, K. C. Morgan.** *Ibid.*, vol. 14, 1958, pp. 1044–57.

13. Binary Communication Feedback Systems, **B. Harris, A. Hauptschein, K. C. Morgan, L. S. Schwartz.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 960–69.

14. Coded Binary Decision-Feedback Systems, **J. J. Metzner, K. C. Morgan.** *Transactions*, Professional Group on Communications Systems, Institute of Radio Engineers, vol. CS-8, no. 2, June 1960, pp. 101–13.

Binary Controls for Error Control

W. W. PETERSON
ASSOCIATE MEMBER AIEE

Summary: The model of a communication channel used in the study of binary codes, its relation to real data transmission channels, and assumptions about noise, will be reviewed briefly. Several error-correcting codes will be listed with a brief description of their error-correcting capabilities and feasibility of their implementation.

The most promising class of codes for implementation at present are cyclic codes. The burst-error-correcting cyclic codes are especially attractive from the point of view of implementation and a simplified block diagram for implementing a practical example of such a code will be shown and discussed.

The Communication Channel[1]

A BLOCK DIAGRAM of a digital communication system is shown in Fig. 1. The same model describes an information storage system, if the storage medium is considered as a channel. A typical transmission channel is a telephone line. A typical storage device is a magnetic tape unit including writing and reading heads. The source information is usually messages of binary or decimal digits or alphabetic information in some form. The encoder transforms these messages into signals acceptable to the channel, typically electric signals with some restrictions on power, bandwidth, and duration. These signals enter the channel and are perturbed by noise. The output enters the decoder, which decides which message was sent and delivers this message to its destination. The communication engineering problem is principally the design of the encoder and the decoder, though it may also include improving the channel. The encoder includes modulation and the decoder includes detection.

The system described in Fig. 1 is too general to be handled effectively in theoretical analyses.

General theory indicates that such a channel has a definite capacity, that typical sources have a definite rate, and that if the rate of the source is less than the channel capacity, it is possible to encode and decode in such a way that the probability of erroneous decoding is small.[2-4] This theory looks promising for future work but it provides no more than vague indications at the present as to how a data transmission system should be designed.

Fig. 2 illustrates a typical present-day system. Information is almost universally handled within a machine in binary form and is converted at the input to binary. Many systems use a code in which various combinations of six binary digits represent each of the numerals, letters, space, and various special symbols such as punctuation. Another common combination is a code using four binary digits for each decimal digit, and two decimal digits for alphabetic or special characters.

The binary-to-channel encoder is sometimes called a modulator. Typically it encodes a "one" into a pulse and a "zero" into either no pulse or one that is readily distinguishable from the code for "one." This handling of a single binary symbol at a time in the channel encoder is a restriction that definitely causes the loss of channel capacity. The channel-to-binary decoder makes a decision for each received pulse as to whether a "one" or a "zero" was transmitted. The independent decoding of individual pulses results in further loss of channel capacity. Theory indicates that more complex encoding and decoding would make transmission possible at higher rates with the same probability of error, but no effective practical way of accomplishing this has been found.

The binary-to-binary encoder and decoder implement binary error-detecting or error-correcting codes.

Some Remarks on Binary Error-Correcting and Error-Detecting Codes

In an ideal system the binary symbol that is received from the channel-to-binary decoder should match the symbol that entered the binary-to-channel encoder. In a practical system there are occasional errors, and it is the purpose of binary codes to detect and possibly correct such errors. These codes cannot correct every conceivable pattern of errors, but must be designed to correct the most likely patterns. Much binary coding theory has been based on the assumption that each symbol is affected independently by the noise, and therefore that the probability of a given error pattern depends only on the number of errors. While this may be appropriate for some channels, on telephone lines and on magnetic tape storage systems errors occur predominantly in bursts. Telephone line disturbances, such as lightning, last longer than the duration of one symbol. Magnetic tape defects are larger than the space required for a symbol. Thus, codes for correcting bursts of errors were required, and some remarkably good codes have been developed. In general, too little is known about the error statistics of typical channels for accurate evaluation of known coding methods and for indicating requirements for practical codes.

The communication channels shown in Figs. 1 and 2 are strictly one-way channels. Frequently communication systems are two-way channels, and this fact must be considered in designing codes. With a two-way channel, for example, an error-detecting code can be used. When an error is detected at one terminal a request for a repeat signal can be given, and in this way errors can be effectively corrected.

There are one-way channels in which error probabilities can be reduced with error-correcting codes but not by error detection and retransmission. With a magnetic tape storage system, for example, a retransmission cannot be requested after the tape has been stored a week or a month and errors are detected when the record is read. Encoding for error-correcting codes is slightly more complex than for error-detection; it is the decoding that is likely to require complex equipment. For transmission from a satellite where the amount of equipment at the transmitter is much

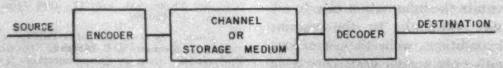


Fig. 1. Block diagram of a general data communication or storage system

---

Paper 61-1097, recommended by the AIEE Communication Theory Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15-20, 1961. Manuscript submitted January 6, 1961; made available for printing August 15, 1961.

W. W. PETERSON is with the University of Florida, Gainesville, Fla.

The author's work in coding was done largely through consulting and joint research with the International Business Machines Research and Development Laboratories and the Massachusetts Institute of Technology Research Laboratory for Electronics, which is supported by the U. S. Army Signal Corps, the U. S. Air Force Office of Scientific Research, Air Research and Development Command, and the U. S. Navy Office of Naval Research. He appreciates his association with colleagues at these laboratories which he feels has been invaluable.

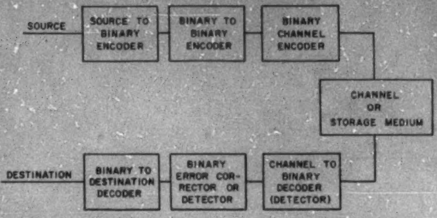

SOURCE
SOURCE TO BINARY ENCODER
BINARY TO BINARY ENCODER
BINARY CHANNEL ENCODER
CHANNEL OR STORAGE MEDIUM
DESTINATION
BINARY TO DESTINATION DECODER
BINARY ERROR CORRECTOR OR DETECTOR
CHANNEL TO BINARY DECODER (DETECTOR)

Fig. 2. Block diagram of a typical data communication or storage system

more important than the total amount of equipment, the encoding equipment for an error-correcting system is likely to be more practical than provision for remotely controlled retransmission. The complex error-correcting procedures can be carried out on earth where equipment limitations are not severe. Thus, there are system problems which call for error-correcting codes requiring no feedback.

On the other hand, error detection is by its nature a much simpler task than error correction. When a two-way channel is available, error detection combined with the provision for a request for retransmission will generally be found more practical than the error-correcting codes known at the present. Some codes, especially the burst-error-correcting codes, discussed in the next section, implement so well that even when there is a provision for retransmission, limited error correction might well be built into the system to reduce the number of retransmissions required.

The efficient use of feedback in error control has not received the attention it deserves in coding theory. Certainly feedback can greatly simplify error correction. Yet there is a limit to the efficiency of a simple error-detection and retransmission system, for short-error detection codes cannot efficiently detect errors, and if extremely long codes are used too much retransmission is necessary. Study of the use of the feedback channel in a more sophisticated way has barely been started.[5,6]

In equipment in use at the present, the binary-to-binary encoder may frequently not be used at all, or some simple error-detecting code such as a two out of five, four out of eight, or simple parity check code might be employed. Recent developments, both in digital computer technology and in the theory of coding, make more powerful error-correcting and error-detecting equipment feasible. Some of the most promising coding methods are discussed in the next section.

Cyclic Error-Correcting and Error-Detecting Codes

The most attractive new error-correcting and error-detecting codes are examples of cyclic codes.[7–10] Cyclic codes are systematic group codes;[11] that is, blocks of k binary information symbols are coded into blocks of n binary symbols by appending $n-k$ parity check symbols. (A parity check symbol is an odd-even check on a subset of the information symbols.) Cyclic codes have additional structure that allows encoding and parity check calculation using shift registers with feedback connections.[10]

Though most work on cyclic codes has dealt with error correction, cyclic codes are generally well suited for error detection. Encoding can always be accomplished by a shift register of the type described in the next section for a Fire code. A check for parity failures at the receiver is made with an identical device. If the shift register does not contain all zeros after the received block has entered it, an error has been detected.

Any cyclic code is able to detect any single burst of length equal to, or less than the number n-k of parity checks.[8] In addition, a code capable of correcting a single burst of length b can detect any combination of two such bursts. A code capable of correcting any combination of t errors can detect any combination of $2t$ errors. Thus, for example, the Fire code described in the next section can detect any single burst of length 16, or less, any two bursts of length 5, or less. (Actually it can be shown that this code is able to detect any pair of bursts provided one has length 5 or less and the sum of their lengths is no greater than 12.) The 5-error-correcting Bose-Chaudhuri code given as an example in a later section has length 1,023 and requires 50 parity check symbols. It can detect any single burst of length 50 or less or any random arrangement of 10 errors. Of course, many other error patterns will also be detected; if a code has n-k parity check symbols, then there is only one chance in 2^{n-k} that a randomly chosen pattern of errors will satisfy the parity checks and result in an undetected error.

Fire Codes[10,11]

Fire has found a class of cyclic codes which correct or detect a single burst of errors in a coded block. A code of any length can be found which corrects any single burst of length b or less and detects any single burst of length $d \leq b$. An arbitrary parameter m relates the length n of the code and the number n-k of parity check symbols. The total length of the code must be no greater than the least common multiple of 2^m-1 and $b+d-1$. The number of check symbols is $m+b+d-1$. For example, if $m=5$, $b=5$, and $d=7$, then a code is provided that is capable of correcting any single burst of length 5 and simultaneously detecting any single burst of length 7. The length of the code is $(2^5-1)\cdot(5+7-1)=341$. There are $5+5+7-1=16$ parity checks and $341-16=325$ information symbols.

The equipment required for encoding is shown in Fig. 3. The square blocks represent single stages of shift register, and the circles represent exclusive or logic blocks. The device operates as follows:

1. With the gate turned on, the information symbols are simultaneously trans-

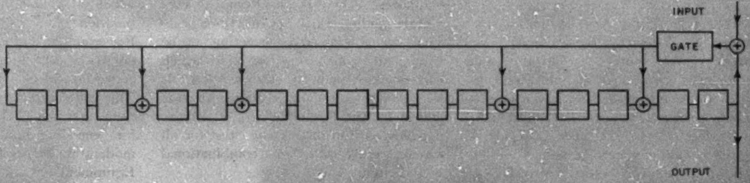


Fig. 3. Encoder for a typical Fire code

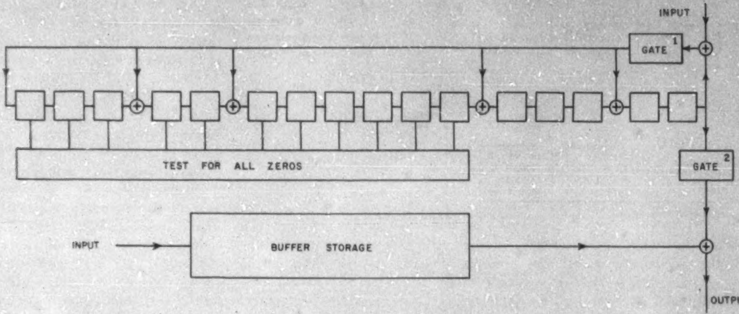


Fig. 4. An error corrector for a typical Fire code

mitted over the channel and entered at the input to the shift register encoder. The shift register is shifted right each time a symbol enters, but the output of the shift register is not allowed to enter the channel.

2. When all the information symbols have entered the shift register and channel, the required $n-k=16$ parity check symbols remain in the register. Then the gate is turned off and these symbols are shifted out into the channel behind the information symbols to make a complete coded block.

The equipment required for error correction is shown in Fig. 4.

The device operates as follows:

1. The entire received block is simultaneously entered into the shift register, which calculates parity checks, and the buffer storage, which simply stores the received block pending the outcome of the parity check calculation. The shift register is shifted each time a symbol enters. Gate 1 is on and gate 2 off during this phase of the procedure.
2. Next the symbols are read out of the buffer one at a time, and simultaneously the shift register is shifted with no input, once for each symbol read out of the buffer. Gate 1 remains on and gate 2 off.
3. As soon as the first 11 stages of the shift register are found to contain all zeros, the error-burst pattern appears in the last five stages, and the next five symbols to come out of the buffer contain the errors. Correction is achieved by turning off gate 1, turning on gate 2, and continuing to read out of the buffer and shift.
4. If the test for all zeros is never met, then an uncorrectable error has been detected.

There are some cyclic burst-error-correcting codes that require even fewer parity check symbols to do the same amount of error correcting, but Fire codes are the only presently known infinite class for correcting more than double-adjacent errors.[13–15] They can all be implemented in the same manner as that shown in Figs. 3 and 4 for a Fire code.

Hamming Codes[16]

For any m there is a code of length 2^m-1 which has m parity check symbols and 2^m-1-m information symbols and is capable of correcting any single error. One parity check can be added to permit double-error detection as well as single-error correction, or alternatively to correct single and double adjacent errors.[17] The symbols in Hamming codes can be arranged to make them cyclic codes.[10,17] They are best implemented as a special case of a cyclic burst-error-correcting code; the equipment required is essentially the same as that described for a Fire code.

Bose-Chaudhuri Codes[10, 18–21]

For any m and t there is a code of length $n=2^m-1$ that corrects all combinations of t or fewer errors. The number of parity check symbols is never greater than mt. For example, if $m=10$, then $t=5$ produces a 5-error correcting code of length $n=1{,}023$, which requires 50 parity check symbols and therefore has $1{,}023-50=973$ information symbols. These are cyclic codes and therefore the encoding and parity check calculation can be accomplished[1] in the same manner as for the Fire code. The error correction is more difficult, however, but feasible. In general, it requires a small special-purpose computer. There is an alternative correction method which appears to be practical for double-error-correcting cyclic codes, but probably not for codes which correct more than two errors.[10,22] It is very similar to the arrangement which is shown for the Fire code, and it is more complex only in having the "test for all zeros" replaced by a combinational logical circuit.

Other Promising Binary Codes

FIXED COUNT CODES

Codes in which all coded blocks contain the same number of ones are used for error detection in several existing systems. For example, a four-out-of-eight code is used in the IBM (International Business Machines) data transceiver.[23] An attractive feature of these codes is that the only undetectable errors are those in which the number of erroneous "ones" equals the number of erroneous "zeros," and these combinations would seem to be unlikely in practical systems.

SIMPLE PARITY CHECK CODES

A single parity check appended to a block of information symbols makes single error detection possible. More error detection, and even error correction, can be achieved by arranging the information in a rectangular array and adding parity checks to rows, columns, or even diagonals. An excellent discussion of such codes is given by Kautz.[24]

REED-MULLER CODES[25,26]

For any choice of m and r, $(r<m)$ there is a Reed-Muller code with coded words of length $n=2^m$ with

$$k=\sum_{i=0}^{n}\binom{n}{i}\text{ information symbols}$$

which is capable of correcting any combination of $2^{m-r-1}-1$ or fewer errors. For example, for $n=7$ and $r=3$, blocks of 64 binary information symbols are coded into blocks of 128 binary symbols. This code corrects any combination of 7 or fewer errors. The equipment required for encoding and error correction is moderately complicated but feasible. Equipment for the previously mentioned

Table I. Examples of Hagelbarger's Codes

Rate (percentage of information symbols)	50	67	80
Maximum length of correctable burst	6	6	10
Necessary error-free guard space between bursts	19	50	194
Shift register stages: encoder	7	26	132
decoder	16	42	168

128 symbol code has been built at Lincoln Laboratory.[27]

It is interesting to compare the Bose-Chaudhuri 7-error-correcting code of length 127 with the Reed-Muller 7-error-correcting code of length 128. The Bose-Chaudhuri code would require only 49 parity check symbols, and therefore would have 78 information symbols, as compared with 64 for the Reed-Muller code. The encoder for the Bose-Chaudhuri code would be considerably simpler than the encoder for the Reed-Muller code. The error-correcting circuitry, however, would be considerably more complicated for the Bose-Chaudhuri code. For longer codes, the Bose-Chaudhuri codes would be expected to be more suitable.

LOW-DENSITY PARITY CHECK CODES[28]

Gallagher has shown that it is possible to construct good low-density parity-check codes, that is, codes for which each parity check rule checks only a few symbols in the coded blocks. He simulated an error-correction procedure for such codes on an IBM *704* computer and found that with codes of length $n=500$, of which half the symbols are information symbols, his procedure was able to correct most patterns of up to about 40 errors. With only one-fourth of the 500 symbols information symbols, most patterns of about 80 errors could be corrected. These results are fairly promising and indicate the possibility of finding rather general coding methods, which do not depend on elaborate algebraic or geometric structures.

RECURRENT CODES[10, 29–31]

A recurrent code is a type of nonblock code in which check symbols are inserted at intervals among the information symbols. Hagelbarger and Kilmer have found recurrent codes that are particularly well adapted to burst-error correction. The encoder and the decoder are constructed from shift registers. Three examples taken from Hagelbarger's paper are shown in Table I.

In general, a Fire code that corrects bursts of the same length as a particular Hagelbarger's code and has the same rate requires a smaller error-free guard space between bursts. To put it another way, a Fire code that corrects bursts of the same length and requires the same guard space has a higher rate. Hagelbarger codes appear to implement more economically than Fire codes if the rate is less than about 75% but less economically for high rates. Some equipment saving over Hagelbarger's estimates are possible with minor modifications of his coding methods.[10]

SEQUENTIAL DECODING OF RECURRENT CODES[32]

Wozencraft has devised a method of decoding for recurrent codes which is well suited for correction of random errors. The encoding requires a simple shift special purpose computer. Simulation experiments with essentially random shift-register connections have been encouraging.[33]

AN + B CODES[34]

Error detection or error correction can be achieved by encoding the number N into $AN+B$ and representing the result in binary form. The amount of redundancy required for a single-error-correcting $AN+B$ code is typically one more binary digit than is required for a Hamming code, but $AN+B$ codes are capable of correcting ordinary single errors and errors equivalent to a carrying failure. Since the encoding and decoding operations are ordinary arithmetic procedures they can be done by any general purpose computer. Thus these codes can be used for data transmission between computers with no special equipment requirements. They can also be used for checking an adder; in fact, a parallel binary adder using one of these codes can easily be made single-component-failure correcting.

Conclusion

This discussion of certain aspects of binary coding together with a discussion of some of the best error-detecting and correcting codes is intended to orient the researcher interested in error control in data transmission or storage systems and to direct him to more detailed technical papers in this area.

References

1. The Transmission of Information (book), **R. M. Fano.** The Technology Press, Cambridge, Mass., 1961.

2. The Mathematical Theory of Communication (book), **C. E. Shannon, W. W. Weaver.** University of Illinois Press, Urbana, Ill., 1949.

3. Probability of Error for Optimal Codes in a Gaussian Channel, **C. E. Shannon.** *Bell System Technical Journal*, New York, N. Y., vol. 38, May 1959, pp. 611–56.

4. A Class of Codes for Signaling on a Noisy Continuous Channel, **J. L. Kelly, Jr.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-6, no. 1, Mar. 1960, pp. 22–24.

5. Coding for Two-Way Channels, **M. Horstein, J. M. Wozencraft.** *4th London Symposium on Information Theory*, Butterworth Scientific Publications, London, England, 1960.

6. Sequential Transmission of Digital Information with Feedback, **M. Horstein.** *Ph.D. Dissertation*, Massachusetts Institute of Technology, Cambridge, Mass., 1960.

7. Cyclic Error-Correcting Codes in Two Symbols, **E. Prange.** *Technical Note AFCRC-TN-57-103*, Air Force Cambridge Research Center, Bedford, Mass., Sept. 1957.

8. Some Cyclic Error-Correcting Codes with Simple Decoding Algorithms, **E. Prange.** *Technical Note AFCRC-TN-58-156*, Air Force Cambridge Research Center, Apr. 1958.

9. The Role of Coset Equivalence in the Analysis and Decoding of Group Codes, **E. Prange.** *Technical Note AFCRC-TR-59-164*, Air Force Cambridge Research Center, June 1959.

10. Error Correcting Codes (book), **W. W. Peterson.** The Technology Press, 1961.

11. A Class of Binary Signaling Alphabets, **D. Slepian.** *Bell System Technical Journal*, vol. 35, Jan. 1956, pp. 204–34.

12. A Class of Multiple-Error-Correcting Binary Codes for Non-Independent Errors, **P. Fire.** *Report RSL-E-2*, Sylvania Electric Products, Inc., Mt. View, Calif., Mar. 1959.

13. A New Group of Codes for Correction of Dependent Errors in Data Transmission, **C. M. Melas.** *IBM Journal of Research and Development*, New York, N. Y., vol. 4, Jan. 1960, pp. 58–65.

14. Codes for Correction of "Clustered" Errors, **S. H. Reiger.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, vol. IT-6, no. 1, Mar. 1960, pp. 16–21.

15. Error-Correcting Codes from Linear Sequential Circuits, **N. Abramson.** *4th London Symposium on Information Theory*, Butterworth Scientific Publications, 1960.

16. Error Detecting and Error Correcting Codes, **R. W. Hamming.** *Bell System Technical Journal*, vol. 29, 1950, pp. 147–50; also *Bell Laboratories Record*, New York, N. Y., vol. 28, 1950, pp. 193–98.

17. A Class of Systematic Codes for Non-Independent Errors, **N. M. Abramson.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, vol. IT-5, no. 4, Dec. 1959, pp. 150–57.

18. A Class of Error-Correcting Binary Group Codes, **R. C. Bose, D. K. Ray-Chaudhuri.** *Information and Control*, Academic Press, New York, N. Y., vol. 3, Mar. 1960, pp. 68–79.

19. Further Results on Error Correcting Binary Group Codes, **R. C. Bose, D. K. Ray-Chaudhuri.** *Ibid.*, pp. 279–90.

20. Encoding and Error-Correction Procedures for the Bose-Chaudhuri Codes, **W. W. Peterson.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, Sept. 1960.

21. Two-Error Correcting Bose-Chaudhuri Codes Are Quasi-Perfect, **D. Gorenstein, W. W. Peterson, N. Zierler.** *Information and Control*, Academic Press, vol. 3, Mar. 1960, pp. 291–94.

22. Error-Correcting Codes for Correcting Bursts of Errors, **J. E. Meggitt.** *IBM Journal of Research and Development*, vol. 4, July 1960, pp. 329–34.

23. A Data Transmission Machine, **C. R. Doty, L. A. Tate.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 75, Nov. 1956, pp. 600–03.

24. A Class of Multiple-Error-Correcting Codes for Data Transmission and Recording, **W. H. Kautz.** *Technical Report no. 5, Project no. 2124*, Stanford Research Institute, Stanford University, Stanford, Calif.

25. Application of Boolean Algebra to Switching Circuit Design and to Error De-

tection, **D. E. Muller.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, vol. EC-3, Sept. 1954, pp. 6–12.

26. A Class of Multiple-Error-Correcting Coding and Decoding Schemes, **I. S. Reed.** *Ibid.*, Professional Group on Information Theory, vol. IT-4, Sept. 1954, pp. 38–49.

27. An Error-Correcting Encoder and Decoder for Phone Line Data, **K. Perry.** *WESCON Convention Record*, Institute of Radio Engineers, pt. 4, 1959, pp. 21–26.

28. Low Density Parity Check Codes, **R. Gallaher.** *Ph.D. Dissertation*, Massachusetts Institute of Technology, 1960.

29. Recurrent Codes: Easily Mechanized, Burst Correcting, and Binary Codes, **D. W. Hagelbarger.** *Bell System Technical Journal*, vol. 38, July 1959, pp. 969–84.

30. Some Results on Linear-Recurrent Binary Burst-Correcting Codes, **W. L. Kilmer.** *Technical Report no. 3, Project no. 6795-102*, Electronics Research Laboratory, Montana State College, Bozeman, Mont., Jan. 1959.

31. Linear-Recurrent Binary Error-Correcting Codes for Memoryless Channels, **W. L. Kilmer.** *Technical Report no. 4, Project no. 6795-102*, Electronics Research Laboratory, Montana State College, Jan. 1960.

32. Sequential Decoding for Reliable Communication, **J. M. Wozencraft.** *Convention Record*, Institute of Radio Engineers, vol. 5, pt. 2, Mar. 1957, pp. 11–25; also *Report no. 325*, Research Laboratory for Electronics, Massachusetts Institute of Technology, Aug. 1957.

33. An Experimental Study of Sequential Decoding for the Binary Symmetric Channel, **M. Horstein.** *Group Report 34-74*, Lincoln Laboratory, Massachusetts Institute of Technology.

34. Error Detecting and Correcting Codes for Arithmetic Operations, **D. T. Brown.** *Transactions*, Professional Group on Electronic Computers, Institute of Radio Engineers, vol. EC-9, no. 3, Sept. 1960, pp. 333–37.

◆

Discussion

J. J. Metzner and **K. C. Morgan** (New York University, New York, N. Y.): In discussing the use of feedback with coding, it is stated in the paper that "The efficient use of feedback in error control has not received the attention it deserves in coding theory . . . Study of the use of the feedback channel in any more sophisticated way has barely been started."

We are in agreement with the first statement in this quotation. With respect to the second, however, we believe that the work done on feedback systems at New York University has been overlooked and these references should be included.

References

1. Theory of Information-Feedback Systems, **S. S. L. Chang.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-2, no. 3, Sept. 1956, pp. 29–40.

2. Optimum Decision-Feedback Systems, **B. Harris, A. Hauptschein, L. S. Schwartz.** *National Convention Record*, Institute of Radio Engineers, pt. 2, Mar. 1957, pp. 3–10.

3. Binary Decision-Feedback Systems for Maintaining Reliability under Conditions of Varying Signal Strength, **B. Harris, A. Hauptschein, K. C. Morgan, L. S. Schwartz.** *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 13, Oct. 1957, pp. 126–40.

4. Binary Symmetric Decision Feedback Systems, **Bernard Harris, K. C. Morgan.** *AIEE Transactions*, pt. I *(Communication and Electronics)*, vol. 77, Sept. 1958, pp. 436–43.

5. Cumulative Binary Decision Feedback Systems, **S. S. L. Chang, B. Harris, K. C. Morgan.** *Proceedings*, National Electronics Conference, vol. 14, Oct. 1958, pp. 1044–57.

6. Binary Relay Communication and Decision Feedback, **J. J. Metzner.** *National Convention Record*, Institute of Radio Engineers, pt. 4, Mar. 1959, pp. 112–22.

7. Reliable Fail-Safe Binary Communication, **J. J. Metzner, K. C. Morgan.** *WESCON Convention Record*, Institute of Radio Engineers, pt. 5, Aug. 1960, pp. 192–206.

8. Coded Feedback Communication Systems, **J. J. Metzner, K. C. Morgan.** *AIEE Special Publication T-123*, 1960, pp. 84–94; also *Proceedings*, National Electronics Conference, vol. 16, Oct. 1960, pp. 250–57.

W. W. Peterson: Dr. Metzner and Dr. Morgan are quite correct. I most certainly should have mentioned and listed the important and extensive work on communications systems with feedback done at New York University. It was not intentionally left out. However, I still feel the surface has barely been scratched in this area.

A 1,650-Bit-Per-Second Data System for Use Over the Switched Telephone Network

S. BRAND
MEMBER AIEE

C. W. CARTER
MEMBER AIEE

Summary: The engineering of a data system for use on the switched message network is reviewed. First, the conditions imposed by this specific application are shown to influence the selection of signal characteristics, and the choice of raised-cosine pulse shaping is given support by a brief treatment of some pulse transmission fundamentals. Finally, the results of a field trial are summarized to show that, while the system objectives were met, factors such as delay distortion, attenuation distortion, and transmission loss limited the performance obtained.

SYSTEMS DEPENDENT ON the transmission of pulse trains between terminals are being developed extensively as the advantages of data and computer techniques become more attractive in various lines of endeavor. Much interest has been shown in the possible use of telephone lines as the transmission medium. Such use can introduce serious problems, particularly if an attempt is made to utilize the communications channel at anywhere near its full efficiency.

The development at Bell Laboratories of a particular data system to work over the Bell System switched network of commercial message circuits with a bit rate of 1,650 bits per second encountered several transmission problems. The solution of these problems and the experience gained in testing the system operation in the plant, although applicable specifically to this system, may have general interest for the whole data field. The principles involved are fundamental in many respects, and emphasize the general problem of pulse transmission over lines and the particular characteristics of telephone plants that require practical consideration from the standpoint of transmission.

The data system described was designed for one specified application and is not contemplated as a standard Bell System offering. Its performance over a relatively small sample of telephone switched plant facilities should not be taken to indicate the ultimate capability of the Bell System to transmit high-speed

---

Paper 61-1063, recommended by the AIEE Data Communication Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted March 1, 1961; made available for printing August 25, 1961.

S. Brand (deceased) and C. W. Carter (retired) were with Bell Telephone Laboratories in New York, N. Y., and Murray Hill, N. J.

The work reported herein was undertaken as part of a project sponsored by the Department of Defense at Bell Laboratories. The participants are too numerous to list, save for these individuals who co-operated during the actual field tests: G. J. Barna, R. D. Fracassi, E. H. Gilson, R. A. Gibby, V. J. Hawks, A. W. Klibbe, and E. C. Thompson. Those who developed the terminal equipment deserve much credit for making the tests possible, and many people in the operating companies of the Bell System contributed greatly in establishing the calls and tracing them to determine difficulties when they were encountered.

data. An evaluation of present capability is given in a paper[1] which is based on a 1,200-bit-per-second system. Newer developments, especially in the field of transmission methods, promise further improved performance.

In this paper, the system features involving transmission engineering are considered, along with certain basic principles from which decisions were made. This is followed by a report on limited field experience with the system.

System Features

Of concern here are those parts of the data terminals required (1) to treat the input random rectangular pulses for proper transmission and (2) to restore the received data signal to its original rectangular form after transmission. These parts are referred to as the line circuits of the terminals. Also of concern is the transmission of these treated or modified signals over telephone circuits between the transmitter and receiver. The overall objective is to accomplish this with the least chance of error.

Certain general features which characterize this data system need to be explained first because they have a direct bearing on some transmission engineering aspects of the design of the circuits. To meet certain requirements regarding system operation, some broad engineering decisions needed to be made early in the project so that development could proceed on schedule. If this development were undertaken today, some of the decisions might be altered somewhat, although the general principles still apply.

The data system is intended for the interchange of information on a to-and-fro basis between two terminals. Each period of transmission may vary in length from a few seconds up to many minutes. There is an analogy to back-and-forth communication by teletypewriter. The data system is operated half-duplex, in one direction at a time. The data terminal is normally in the RECEIVE condition and in order to SEND, a transmit key must be operated.

It is an error-checking system in a sense for certain short intervals of time. At the start of each period of transmission, a special train of pulses is sent to the receiving terminal. There the terminal regains synchronism from the signal and is conditioned automatically for reception. There are internal features which check to see whether this has been done successfully. At times, because of noise or some other adverse line condition, the operation may not be correct. In such a case, the receiving terminal sends an alarm tone back to the transmitting terminal to advise the user there to release the transmit key and try again. If successful this time, no alarm is returned.

No Adjustment

With this system, the user is not expected to make adjustments. He is not to be a technical operator. The absence of adjustments, coupled with the high bit rate and our previous experience with the message plant, made it plain from the start that the system should be designed for every bit of transmission margin possible.

Line Signal—Bit Rate

The bit rate of 1,650 bits per second was dictated by circumstances which need not be considered here. This rate is about as low as can be tolerated for the purpose at hand, and is about as high as was considered practical at the time for the message band from preliminary line tests.

The principal transmission features of the terminals are:

1. Two levels only—no framing.
2. Vestigial sideband (VSB), or amplitude modulation. Carrier—2,200 cps; modulation—50%.
3. Line signal approximately 700 cps to 2,600 cps.
4. Automatic gain control (AGC) range—35 db (decibels).

These now are discussed briefly.

Binary Pulse

Since no special timing or framing pulses were needed, a simple binary pulse system could be used. This was an advantage because this sort of signal is more tolerant of noise than a multilevel signal.

Vestigial Sideband

As the basic system signal cannot be transmitted at baseband frequency in the telephone plant, some type of carrier technique was necessary. Earlier investigations indicated amplitude modulation as holding most promise among well-understood methods at the time. The choice of VSB, rather than Double Sideband (DSB), resulted from the taboo on adjustment. Preliminary tests showed that DSB was practical on fairly long message circuits, even at this high bit rate.

But it was also found that with circuits such as *K*, *L*, and *R* (*Type A* channel banks), the carrier should be at about 1,800 cps. The *K*, *L*, and *R* are broadband carrier systems using frequency division terminal equipment (channel banks) to derive many 3-kc message channels per system. *K* is a system for nonloaded paired cable, *L* is for coaxial cable, and *R* is for radio relay. The transmission characteristics of these 3-kc channels are determined largely by the multiplex equipment.

On *H44* loaded lines, the carrier had to be about 1,200 cps. *H44* lines are paired cable with 44-millihenry loads on 6,000-foot spacing with voice-frequency repeaters about 45 miles apart.

The delay distortion on these circuits is such that, for a compromise carrier frequency of 1,500 cps, the lower sideband is distorted too much on carrier lines and the upper sideband on loaded lines. There is no suitable compromise frequency, so VSB was adopted.

VSB's disadvantage is that a quadrature component is created, which increases the intersymbol interference, but this undesirable effect can be reduced by lowering the modulation depth. As the modulation lowers, the signal becomes less tolerant to interference from noise. The two considerations play against each other in such a way that there is an optimum, and this is somewhat below 50% modulation. The meaning of 50% modulation is that, if the signal is all marks, the value is unity; then, if the signal is all spaces, the value is one-third. The value selected for this system was 50%.

Choice of 2,200-CPS Carrier

A choice had to be made between full upper sideband on a carrier of about 1,200 cps, or lower sideband on a carrier in the region of 2,200 cps. There was reason to believe that 2,200 cps would be better with respect to Kendall Effect (concerning which there is more later) and that 1,200 cps would work better on loaded lines. Tests were made with each value, but there was no sharply conclusive evidence.

The selection of the carrier frequency in the end was dictated by a consideration in connection with the error-checking alarm tone alluded to earlier. A question arises regarding the return of the alarm tone when there is an echo suppressor in the circuit held operated by the data signal. This suppressor is a voice-operated device for use on 4-wire circuits, which disables one direction of transmission when the other direction is in use. This prevents the transmission of echoes on the circuit resulting from imperfect balance at the 4-wire–2-wire terminals of the circuit.

The alarm tone can be returned if the data-signal carrier is at a high frequency

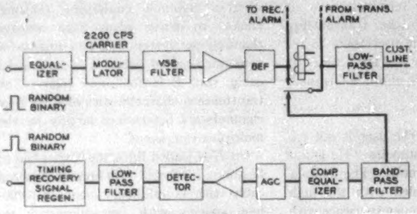


Fig. 1. Transmission elements of system

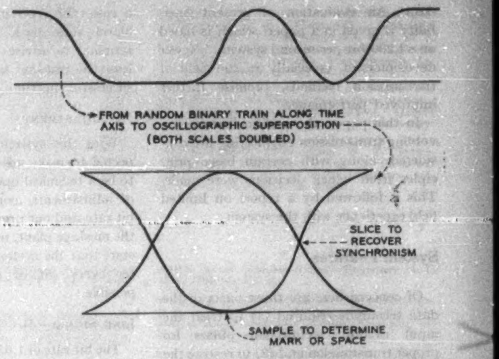

FROM RANDOM BINARY TRAIN ALONG TIME AXIS TO OSCILLOGRAPHIC SUPERPOSITION
(BOTH SCALES DOUBLED)
SLICE TO RECOVER SYNCHRONISM
SAMPLE TO DETERMINE MARK OR SPACE

Fig. 2 (right). Idealized eye picture

in the message band and the alarm tone is near 1,000 cps. This works out because: (1) the *1A* echo suppressor has differential action to permit the louder of two talkers to break through, and (2) the control circuit of the suppressor is frequency-shaped to protect against noise, so that it is some 15 db more sensitive to 1,000 cps than to 2,000 cps.

As a result of tests, the decision was made to use 2,200 cps for the carrier and a component of 1,100 cps in the alarm signal. These particular frequencies were chosen because of their relation to the bit rate of 1,650 per second and their ready availability in the terminal.

The remaining factor to be considered was the effect of this choice on existing signaling systems. In the Bell System plant, single-frequency signaling systems are based on signal frequencies of 1,600; 2,000; 2,400; or 2,600 cycles. This equipment applies a signaling tone to the circuit in the IDLE condition and removes the tone in the TALKING condition. In the latter condition, the signaling receiver is conditioned to accept signaling tone from the distant end (resulting from an ON-HOOK signal) to disconnect the circuit, but it is guarded against false operation by speech on the circuit.

Occasionally, speech with high energy at the signaling frequency may give false operation in spite of the guard, and this effect is known as "talkoff." Signals other than speech, applied to a circuit equipped with these signaling units, may also cause talkoff, and it is essential that data signals do not do this. Tests were made in the laboratory with random data signals on a 2,200-cycle carrier with the four types of signaling equipment already mentioned. The results indicated no tendency to talkoff in any of these cases.

Transmitter Output Power

The actual level used in any particular case depends on local conditions. As explained later, the data terminal is connected to the long-distance switched network through a special subscriber line. The level used on this facility should be as high as possible without causing overloading or interference into other channels on the same route. The main concern is the delivery of a common level to the switched network, high enough to get the best margin against noise and other interferences in the plant without causing undue disturbance to other channels in the same or other systems. Here, consideration must be given to all possible circuits encountered in the telephone network, along with the changes in level resulting at switching points.

Automatic Gain Control

When the terminal is in the RECEIVING condition, there is an AGC having a range of about 35 db. By adjustment of the receiver gain, this range can be made effective for different levels to give a minimum sensitivity threshold of −33 dbm (power in db referred to 1 milliwatt) for a steady mark. Probability studies made early in the development indicated that this range, and sensitivity, with the allowable transmitting level, would be adequate on all but a small fraction of Bell System connections. More will be said about this when discussing the results of the field trial.

The AGC action is such that it maintains the peak of the received signal at a constant level within its operating range. For a distorted signal with large overshoots of certain pulses, this lowers the detected baseband signal relative to that for undistorted signals and requires a corresponding lowering in the slice level. This latter lowering recovers information for regenerating the signal, which would result in the best general performance. This is done at the expense of margin for the undistorted case, where the margin is normally the best.

Time constants of the AGC are rather special. A conditioning period at the start of each transmission has already been mentioned. At the start, the attack time is about 6 ms (milliseconds), but it is slowed down to about 40 ms during the period of signal transmission, after synchronism has been regained. This was found to reduce materially the multiple errors from a single impulse of noise. The recovery time is quite slow, of the order of 3 or 4 seconds for full range.

Transmission Elements

The transmission elements of the terminal are shown in Fig. 1. As already pointed out, the requirement is to accept message-bearing rectangular pulses, to get them over the Bell System message network, and to deliver an accurate replica of them. Some of the features indicated in Fig. 1 have already been covered briefly. Two others are considered more thoroughly since much attention was devoted to them. One is the shaping of the signal elements according to Nyquist's precepts.[2] The other is the compromise delay equalizer. Before taking up the matter of signal shaping in the terminal, some fundamentals of pulse transmission are discussed.

Pulse Transmission

The basic concepts of pulse shaping were given in Nyquist's paper. Nyquist showed that transmission by rectangular waves is remarkably redundant in that all of the message-bearing information is contained in the frequency band from zero to the speed of signaling (half the bit rate), and is then repeated in bands of

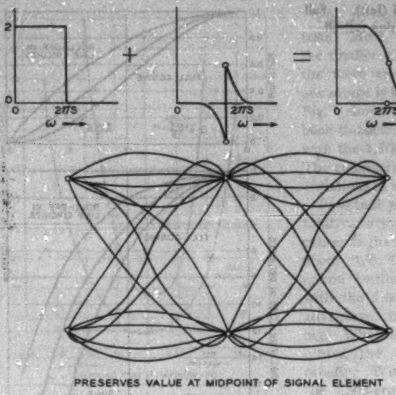


Fig. 3. Nyquist criterion 1

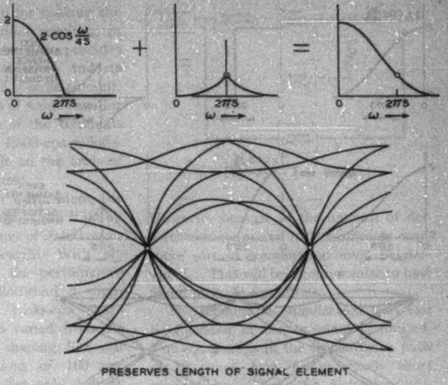


Fig. 4. Nyquist criterion 2

this width all the way to infinity. This indicates that the required information need not be lost in a practical system with a finite cutoff. Theoretically, the transmission path can be restricted to sharp cutoff at the speed of signaling, but this is associated with infinite delay. Practically, a wider band is necessary, and the problem is to determine what cutoff to use, both in regard to the upper frequency limit and the shape of the attenuation curve or rolloff.

In discussing Nyquist's criteria of shaping and their relation to pulse transmission here, the concept of idealized "eye pictures" is introduced as a helpful tool in explaining the effects. Development of the received eye as obtained at the output of the receiving low-pass filter, or baseband, in this system is taken up first, and is followed by a discussion of the shaping principles of interest here.

RECEIVED EYE

The upper part of Fig. 2 represents a few bits of a train of shaped random binary pulses on a horizontal time axis as it might be seen on the screen of an oscilloscope for a single sweep. If the oscilloscope is now synchronized with the bit rate and the sweep adjusted to cover two bits in the display, the separate bits of the train will be superposed to give a picture showing all possible transitions from mark to space and vice versa as seen at the bottom of Fig. 2.

The center opening in this picture, which looks somewhat like an eye, is of considerable interest in data performance since it represents the condition in a pulse receiver for distortionless transmission of the shaped pulses shown in the upper part of Fig. 2. In an ideal system, the pulses would be sliced and sampled at the critical points indicated to regenerate the signal.

PULSE SHAPING

In transmitting a pulse through a medium with distortion, the pulse shape will be distorted, and it is necessary to recover the original pulse from the distorted pulse in the receiver. The subsequent discussion goes on the assumption that there is no phase distortion, and the effects of attenuation distortion only are considered. For a medium with a definite cutoff, limiting the frequencies transmitted to a certain band as required in practice, the pulse is spread out in time with preceding and following transients which can cause interference with other pulses (intersymbol interference), before or after it in the train. These interferences have varied effects on a particular pulse shape, and their importance depends on the particular pulse feature to be preserved. It is possible so to shape this cutoff or rolloff that certain features may be maintained, regardless of pulse sequence and, in general, this shaping is different for the individual feature under consideration.

Two features of a pulse or signal element considered by Nyquist are:

1. Value of the signal element, at its midpoint.
2. Length of the signal element, between crossovers.

Preserving both value and length is important in this system's operation, mainly for sampling and the recovery of synchronism. To maintain the first, Nyquist gives a criterion, here called no. 1, which can be met by infinitely many shapes, as shown in Fig. 3. The narrowest frequency band possible is at the left in Fig. 3 and has a cutoff at $2\pi s$, where $s = 1/2$ the bit rate. If to this shape is added any shape with *odd* symmetry about the speed of signaling, then the midvalue is preserved. That is, the tails of all the pulses produced by a train of unit pulses—except the desired one—go through *0* at each sampling instant and so cause no interference in this respect. The eye picture at the bottom of Fig. 3 shows this. Meeting this criterion gives no assurance necessarily as to the space between crossovers.

It should be pointed out here that this over-all shaping includes the line-circuit shaping characteristic and the spectrum of the transmitted pulse. For a rectangular pulse occupying the full time slot, the spectrum is a maximum value, multiplied by $\sin x/x$ on a voltage basis, where $x = \pi f/2s$. Thus, in a system which is to transmit rectangular pulses, or boxcars, the actual shaping to be provided in the line circuits in the system, baseband to baseband, is the difference between the shaping shown in Fig. 3 and $\sin x/x$ shaping. That is, on an amplitude basis, the required shaping equals the desired over-all shaping, multiplied by $x/\sin x$.

To maintain the length of the signal element, Nyquist gives another criterion, here called no. 2. The basic shaping for this case is shown at the left in Fig. 4. If to this shaping any curve is added with *even* symmetry around the speed

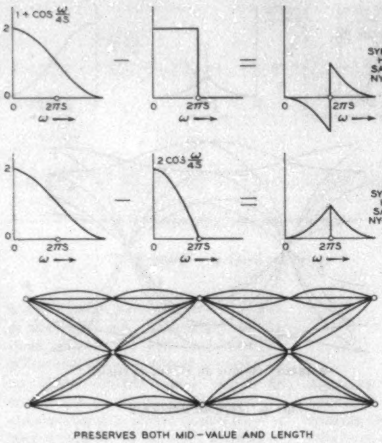

1 + cos ω/4S
2
0
2πS
ω →
—
=
ODD SYMMETRY HENCE SATISFIES NYQUIST I
2 cos ω/4S
EVEN SYMMETRY HENCE SATISFIES NYQUIST II
PRESERVES BOTH MID-VALUE AND LENGTH

Fig. 5 (left). Full cosine roll-off

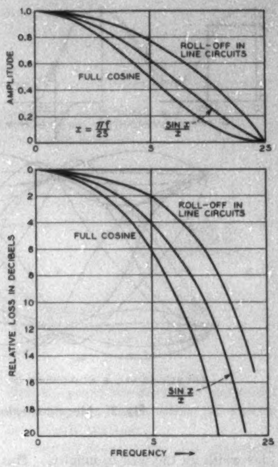


Fig. 6 (right). Roll-off characteristics

of signaling, then the crossover points are maintained as shown by the eye picture at the bottom of Fig. 4. Meeting this criterion gives no assurance necessarily as to the value at midpulse. Here again, the sin x/x shaping factor must be considered.

For this system, it is desirable to meet both criteria, as was stated before. Investigation shows that there is one shaping, and only one, which does this. It is shown in Fig. 5 and consists of a full cosine rolloff from *0* to the full bit rate. It has half of the maximum value (6 db down) at one-half the bit rate. The chart shows graphically how the two criteria are satisfied. The eye picture at the bottom of Fig. 5 illustrates how this rolloff buttons the eye down at the four desired points.

Hence, the ideal added rolloff in a system for rectangular pulses is the full cosine multiplied by $x/\sin x$, or the difference of these shapings on a db basis. This is shown in Fig. 6. Here the maximum amplitude is taken as unity. Actually, the desired shaping is not physically realizable (since it is required to be exactly zero above the bit frequency), but it can be approximated. The shaping used carries the approximation over the 150-cps to 1,500-cps range.

Because of this approximation and particularly because additional amplitude and delay distortion are introduced by the transmission medium, an ideal eye such as the one in Fig. 5 is not fully realized in this system. However, if an area inside the eye can be kept open, and slicing and sampling occur in this area, the signal can be recovered. This is what is attempted in the actual operation of the system.

The location of the final residual area of the eye under conditions of extreme distortion may vary with the cause of the distortion. Then, the slicing and sampling values, being fixed at some particular value, may fall outside the eye. However, compromise settings of slicing and sampling can be determined to fit the most likely location of the residual hole for the limiting values of distortion and interference.

Signal Shaping or Rolloff

Originally it appeared that, to prevent overlap of the baseband spectrum and the lower sideband on the 2,200-cps carrier, the baseband width should be limited to 1,100 cps. Reference should be made to Fig. 7. By so doing, the possibility of Kendall Effect could also be eliminated. Kendall Effect concerns second-order modulation products in the line, which may fall in the transmitted band. If overlap occurs, distortion may result.

Hence, as a start, a 1,100-cps cosine rolloff (6 db down at 825 cps) was provided for test. It was realized that this did not meet criterion no. 2, and that, for a

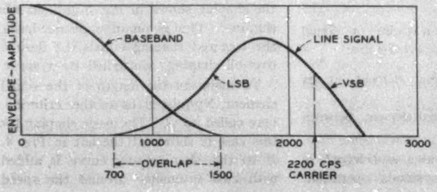


Fig. 7 (left). Baseband and line signal overlap

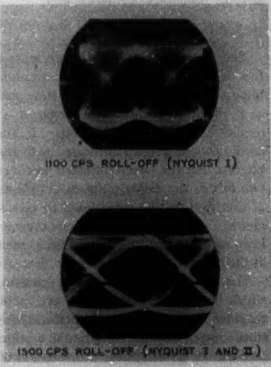

1100 CPS ROLL-OFF (NYQUIST I)
1500 CPS ROLL-OFF (NYQUIST I AND II)

Fig. 8 (right). Pictures of received eyes; terminals connected back-to-back with no compromise equalizer

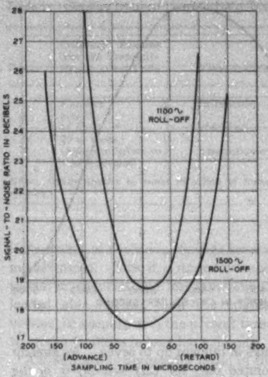


Fig. 9. Effect of sampling time

random signal, there would be sufficient spread of the crossovers to result in horizontal closing of the eye, making sampling difficult. For this reason, a 1,500-cps cosine rolloff was also provided for test. Both criteria are met with this network for all practical purposes. As a result of tests comparing the 1,100-cps and 1,500-cps cosine rolloffs, the latter was finally adopted, as will be explained. It gave much better results without apparently introducing any real trouble from Kendall Effect, and a full 1,650-cps cosine rolloff would probably be chosen for future systems. It should be pointed out that, for the different rolloffs tried, the vestigial cutoffs were not changed.

In an amplitude-modulated system, examination of the received eye is often a good clue in comparing the merits of different conditions. Fig. 8 shows photographs of eye pictures in the obtained receiver with the two rolloffs and the terminals connected back-to-back with no compromise equalizer. Both satisfy Nyquist no. 1, as shown by the openness of the eyes at the sampling points. The 1,100-cps rolloff does not satisfy Nyquist no. 2, as shown by the spreading out of the midpoint crossings. The 1,500-cps rolloff has pulled these together.

The superiority of the wide shaping is shown more clearly in Fig. 9. For quantitative results in comparing different conditions, the test procedure adopted was to introduce noise between the terminals and to determine the signal-to-noise ratio at which the particular error criterion for this system was met for the condition under consideration. (For engineering purposes, a round number of 5 errors per minute was adopted as the criterion.)

Fig. 10 shows a setup for making the tests. Here, the condition resulting in the smaller signal-to-noise ratio provides the more rugged transmission, and its advantage is known by a specific amount. Fig. 9 shows the results of noise tests for back-to-back operation of the terminals with the 1,100-cps and 1,500-cps rolloffs. The comparison is made on the basis of variations in sampling time.

In actual practice for this system, the total variation in sampling time from all causes might cover a range of ±100 ms or about 1/6 the bit length. With accurate center timing, the performance with the 1,500-cps rolloff had a slight advantage over the 1,100-cps shaping. When timing was varied from the center, the 1,100-cps shaping lost out rapidly. With deviations of 100 ms, the 1,500-cps shaping lost only 2 db in ruggedness while the 1,100-cps shaping lost 9 db. The advantage of the 1,500-cps rolloff was even more pronounced when distortions caused by limiting-line transmissions were included.

The foregoing tests were made with thermal noise. Similar results were obtained with impulse noise. As a consequence of these tests, broad shaping was adopted even though overlap is present and there is a possibility of Kendall Effect. Actually, no indication of these effects appeared in any experience in the plant with this system while using the 1,500-cps rolloff. For other pulse data systems, however, Kendall Effect could be serious. For some analog systems, it is known to be a factor to reckon with.

With this over-all rolloff in the terminals, the effective frequency band of the line signal is considered to be from 700 to 2,600 cps.

Compromise Equalization

Up to this point, little consideration has been given to the effect of the lines between the data terminals, but the lines introduce both attenuation and delay distortion. As a result, the data signals are distorted, depending on the sequence of pulses in the train, and the eye tends to close so that, in extreme cases, error-free recovery of the original pulses is difficult. For the intended field of application of this system in the switched network, a great variety of facilities can be encountered in tandem in both the exchange and toll plant.

Because of the requirement that no means of adjustment be available in the terminal for a technical attendant or otherwise, the over-all distortion cannot be equalized for each case as it occurs.

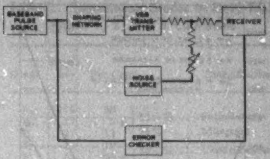


Fig. 10. Test setup

However, because of the amount of distortion encountered in the extreme case, some sort of equalization must be provided. This will be a compromise to best fit all possible conditions.

The amount of equalization that can be included in the terminal is limited. The equalizer is in the transmission path at all times, even for extremely short connections. Thus, the distortion produced by the equalizer itself cannot exceed the allowable amount for a back-to-back connection of the terminals. This means that the line distortion taken care of by this equalizer is also limited. This fact was realized at the start of the project, but it was felt that these restrictions would not cause undue hardship.

To simplify the problem of equalization, advantage is taken of the fact that the line connecting a terminal to the switched network remains unchanged. This connecting line is equalized in itself so that it introduces little distortion in the over-all circuit. This then leaves only the switched long-distance lines to be taken care of by the compromise equalizer in the terminal. The requirements established for the local line are that, over the effective band from 700 to 2,600 cps, the attenuation shall be flat within ±1 db and the envelope delay within ±100 ms.

Before the field trial of this system, it was thought that the attenuation distortion encountered would not be bothersome. However, in a few cases, this belief was proved untrue. Also, on the average, it appeared that there was some excess attenuation at the higher frequencies. Therefore, a "twist" equalizer at line frequency is being included in the terminal, which gives 2- to 3-db correction at 2,200 cps, relative to 1,000 cps.

The matter of delay distortion remains to be accounted for. First, some ideas concerning switched telephone lines are offered. Long connections are built up by putting toll trunks in tandem between switching points. These trunks are often referred to as switching links. However, a toll trunk may consist of several circuit units in tandem, connected together at strategic intermediate points, which are

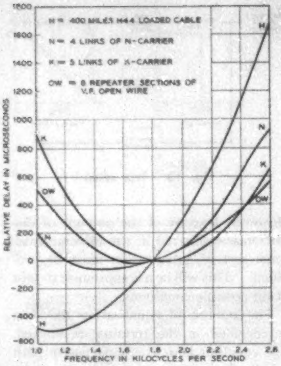


Fig. 11 (left). Envelope delay characteristics of typical message facilities

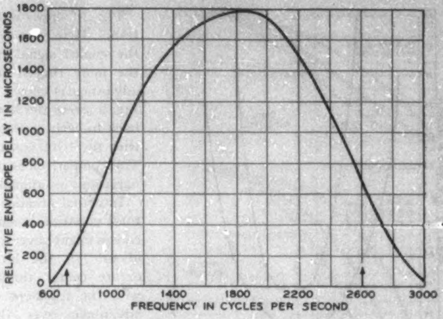


Fig. 12 (right). Compromise delay equalizer characteristic

established for convenience in plant flexibility and maintenance.

At some of these locations, all circuits appear at voice frequency for common usage and, when carrier-derived channels are involved, this means the introduction of terminal channel banks with their associated delay distortion. Thus, a single trunk may include several voice-frequency links. Since the delay distortion in carrier channels is largely caused by terminal equipment, these voice-frequency links in tandem are important here, regardless of the number of switched links in the over-all connection.

Curves which illustrate the principal types of envelope delay distortion in the message plant are shown on Fig. 11, each for substantial lengths of specific facilities. In considering the wide variations among the many possible shapes, and in designing equalizers to fit them, slope-sag diagrams were found helpful.[8]

Two broad categories of telephone lines control the situation in that they have characteristics of a conflicting nature. These are the *K*, *L*, and *R* (*R*=radio) circuits, which use *A* channel banks, on the one hand, and the voice-frequency loaded circuits on the other. (Refer back to explanations under "Vestigial Sideband" section.) The *K*, *L*, and *R* are much more important with respect to circuit mileage, but there are still a good many loaded circuits. The conflict between *K* and loaded lines is in the locations of the minima: 1,800 cps in the first case, 1,000 cps in the second. With more links in tandem, the curves rise more steeply: with fewer links, less steeply; but the minima remain more or less as they are.

A question in connection with the compromise delay equalizer was: How many links in tandem should be equalized? The current-switching plan talks of switched links in tandem up to seven. As already pointed out, for this system, it is the number of voice-to-voice links which matter (since the *A* channel banks largely determine the delay distortion) and there are many single switched links which contain two or even three voice links.

As it is not practical to handle these extreme limits, the compromise shown in Fig. 12 was provided. This equalizer is introduced at line frequency in the terminal. It permits operation with about 7 voice links of *A* channel banks, or about 220 miles of *H44*. It also works satisfactorily back-to-back, but with a substantial degradation in margin.

The envelope delay distortion introduced by the compromise amounts to about 1,500 ms over the signal range. The delay equalizer shown here is not the one originally specified 4 years ago, but

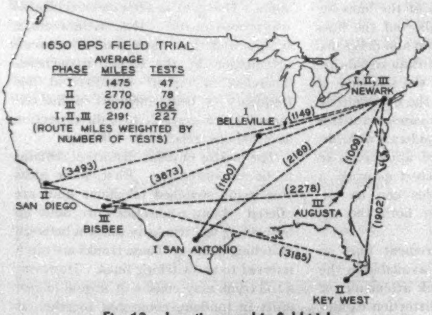

1650 BPS FIELD TRIAL
PHASE
AVERAGE MILES
TESTS
I 1475 47
II 2770 78
III 2070 102
I, II, III 2191 227
(ROUTE MILES WEIGHTED BY NUMBER OF TESTS)
I, II, III NEWARK
I BELLEVILLE
II SAN DIEGO
III BISBEE
I SAN ANTONIO
III AUGUSTA
II KEY WEST
(1149)
(3493)
(3673)
(2169)
(1100)
(1009)
(2278)
(1902)
(3185)

Fig. 13. Locations used in field trial

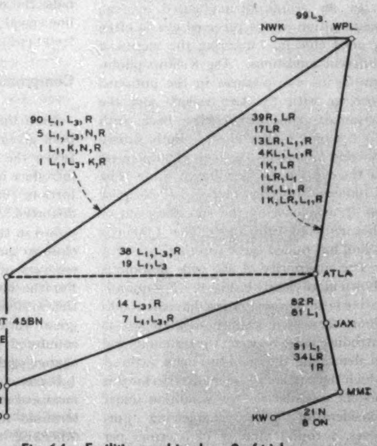

99 L3
NWK
WPL
90 L1, L3, R
5 L1, L3, N, R
1 L1, K, N, R
1 L1, L3, K, R
39R, LR
17LR
13LR, L1, R
4KL, L1, R
1K, LR
1LR, L1
1K, L1, R
1K, LR, L1, R
38 L1, L3, R
19 L1, L3
LA
ATLA
118 L3
99 N
87 LENKURT 45BN
17 VF CABLE
8 K
8 R
14 L3, R
7 L1, L3, R
82R
81 L1
JAX
91 L1
34LR
1R
SBO
127R
SD
MMI
KW
21 N
8 ON

Fig. 14. Facilities used in phase 2 of trial

Table I. Grading Method Used in Field Trial

Grade	Criterion: Failures Per 100 Transmit-Synchronism Tests	Classification
A	3 or less	Satisfactory
B	4 to 10	
C	11 to 25	Marginal
D	26 or more	Unsatisfactory
F	Complete failure to synchronize	

Note: Grade applies to worse direction of transmission.

is the result of evolution through several modifications. Now it is more favorable to carrier channels and less favorable to loaded-cable circuits than the original version, in keeping with the present trend to carrier facilities in long-haul circuits.

Allowance was necessarily made in the equalizer for the delay contributed by many switching-trunk circuits when 5 links are in tandem. These circuits are arrangements of equipment in the office associated with trunks at switchboards or other switching devices for proper supervision, signaling, and similar uses. The delay distortion from them is primarily at the lower frequencies. But, with a number in tandem, the cumulative effect reached into the effective transmission band here.

Field Trial

LOCATIONS, FACILITIES

A field trial of the 1,650-bit system took place in three phases, from mid-February to mid-May of 1958. In each phase, three toll centers, shown in Fig. 13, were involved, one of these always being Newark, which serves the Murray Hill laboratory. Except for the Pacific Northwest, the entire United States was well covered.

The test locations were connected by special toll-subscriber lines directly to a manual position at a toll switchboard in the various centers. As already explained, this had transmission advantages by allowing equalization for delay and attenuation distortion, and control of the loss of these terminal links. These measures virtually eliminated effects other than noise normally encountered in the exchange plant. The data which will be shown shortly are, therefore, largely a picture of performance on long-distance plants only.

One point to be noted is the relation between the connections tested and Bell Systems statistics. The average length of a Bell System toll call (long lines) in 1957 was 233 miles, compared with the average field-trial length of 2,191 miles. In the Bell System, the average number of switched links is two; in the field trial, it was nearly four.

The almost bewildering variety of facilities encountered in these long connections is apparent from Fig. 14 for phase 2 of the tests. This shows first-choice layouts between the three toll centers involved, and also the final choice for San Diego to Key West by way of Los Angeles. The choices refer to alternate routes in the nationwide toll-switching plan. The 77 Atlanta-to-White Plains circuits contain only 3 kinds of facilities: *K*, *L*, and *R*, but they are arranged in 8 different patterns. Some had 4 voice links (a comma means connection at voice). There are group connectors—devices for the interconnection of 12 adjacent channels as a group—in 5 of these patterns. The 337 circuits from Los Angeles to San Bernardino contain 6 kinds, and cover the gamut from voice frequency to L_3 carrier, including Lenkurt and Type *N* carriers, both of which are on nonloaded cable pairs.

GRADING OF PERFORMANCE

In describing the data system, it was mentioned that each period of transmission is initiated by a stream of synchronizing and conditioning pulses. This is a critical point in the operation, and it offers a direct measure of how satisfactorily this data system performs. That is, this period becomes the basis of a criterion of error when the number of successes in every 100 tries is counted. A series of 100 transmit-synchronism tests was adopted as fundamental.

When some experience had been obtained in transmitting data to-and-fro between two users, it was decided to expand this basic idea into a system for grading the performance in terms of the user's satisfaction with the service. This was a useful tool in classifying the field-trial results. The grading system is shown in Table I.

FIELD TEST RESULTS

In the field trial, 12 full-period calls and 215 random calls, totaling 227, were graded. Full-period calls were on a special circuit, set up for test on an over-all 24-hour-a-day basis with no switching. All calls were between 9 pairs of cities. Both success and difficulties were experienced on each of the routes, and the combined grades are shown in Table II.

Remembering that the calls averaged 2,191 miles in length, and the routes varied from 1,009 up to 3,873 miles, the results cannot be considered as picturing the Bell System, but rather a particular network, and a specific type of data system. In brief, on this network hardly more than half the connections were satisfactory, and about a third, numbering 74 connections, were not useful. However, there are reasons for the difficulties experienced, and these are discussed next. As explained later, about 80% of these calls would be satisfactory if repeated now, with appropriate modification of data terminals.

Table II. Transmission Grades in Field Trial, Phases 1, 2, and 3 Combined

Grade	Number	Per Cent		Classification
A	114	50	55+%	Satisfactory
B	12	5+		
C	27	12	12%	Marginal
D	23	10	32+%	Unsatisfactory
F	51	22+		
	227	100		

Causes of Low Grades

Throughout the field trial, emphasis was placed on finding causes for imperfections, and the reasons assigned for failure of the 74 *D* and *F* connections are shown in Table III.

The causes are often not unique. For example, excess delay distortion and excess attenuation distortion frequently occurred together. The one that is believed to be predominant is recorded. Diagnostic accuracy was frequently confirmed by direct action. For example, several grade *F* connections were changed to grade *A* by adding delay correction. *N* overshoot, which will be discussed later, was checked by transferring to circuits known by test to be sufficiently free of overshoot.

Four causes account for the difficulties on 63 out of the 74 calls, as Table III shows. Only two causes had to be labeled

Table III. Field Trial Connections Graded D or F, Causes Assigned

Causes by Group	No.	Totals
Plants designed for voice; not data		
Delay distortion—too many links	24	
N overshoot	12	36
Plants, maintenance		
High loss	16	
Loss distortion—twist	11	
Low loss—singing	1	
Noise or fading	1	
Echo suppressor	1	
Beats or flutter	2	32
Mistakes		
Operator pulled down	1	
Unknown—pulled down during trace	2	3
Data terminals	3	3
	74	74

Table IV. Field Trial Measurements

		1,000-Cps Net Losses of Toll Circuits		
	Route Miles	Minimum	Average	Maximum
Augusta-Newark	1,009	2.1	8.3	22.9
Belleville-San Antonio	1,100	8.3	10.9	24.3
Belleville-Newark	1,149	0.2	8.0	17.8
Key West-Newark	1,902	2.1	9.7	19.4
Newark-San Antonio	2,169	−4.5	5.9	17.2
Augusta-Bisbee	2,278	0.9	7.6	17.9
Key West-San Diego	3,185	−2.0	6.6	15.5
Newark-San Diego	3,493	−3.7	4.7	13.2
Bisbee-Newark	3,873	−8.4	6.1	14.4
All Routes		−8.4	7.5	24.3

unknown; these calls were pulled down by the operator or plant forces before transmission tests could be completed. Difficulty because of delay distortion, high loss, or attenuation distortion was expected, whereas the virtual absence of noise trouble was somewhat surprising.

NOISE

Contrary to expectations, circuit noise proved to be unimportant in the field trial, the only entry on Table III being a case of high noise during a radio fade. Noise on radio circuits was anticipated as a source of trouble, as was impulse noise on the *N* carrier. The latter had been encountered on loops from New York to Stamford and to Hempstead during preliminary tests. Although nothing but *N* facilities were available for getting in and out of Belleville and Key West, noise was no factor in either case. When margins were tested by lowering the transmitting level, the AGC range was exhausted before the signal-to-noise ratio became too small, excepting the one case. The difficulties encountered on the *N* carrier were of another nature—namely, *N* overshoot, discussed later.

DELAY DISTORTION—MULTILINK CONNECTIONS

The compromise delay equalizer designed for the trial was adequate for 5 voice links of facilities, using *A* channel banks (*K*, *L*, and *TD*-2). In the network for the field trial, there were many circuits with more than 5 such links, and some of them were responsible for the unsatisfactory connections.

While a compromise delay equalizer may be designed to cover as many as 7 links of *K*, and also work back-to-back, this severely limits the length of voice-frequency loaded cable. As pointed out earlier, such a modification was made as a result of the trial. Probably no single compromise delay equalizer will handle all connections in the Bell System.

N OVERSHOOT

There is a circuit condition in *N* carrier systems that has little effect on message telephony but, nevertheless, caused this data system some difficulty. The terminals in the *N* carrier systems include compandors for compressing the volume range of the signal on the line to gain noise advantage. A compressor unit is used at the transmitting terminal and an expandor unit in the receiving end. These units have built-in time constants to control the attack and delay times of the compandor action and these should be complementary.

The electrolytic capacitors in the time-constant circuits of the earlier compandors had a tendency to age and, in so doing, rose far above their nominal value of capacitance. This effect no longer occurs where the faulty components in the earlier equipment have been replaced. However, until replacements are completed, some trouble may be encountered with start-stop data systems in which the carrier is interrupted. The effect is to produce overshoots in signal level which cannot be readily handled by AGC circuits.

Since 12 of 31 poor connections at Belleville were ample to establish the *N* overshoot effect, further difficulties of this kind were eliminated in the trial. A few preliminary measurements at Key West, where the only way out was *N* or *ON* carrier, indicated the probability of data transmission difficulty; so all of the time-constant capacitors on the Key West to Miami trunk group were changed. As a result, there was no difficulty with *N* overshoot on connections from Key West.

HIGH NET LOSS

Sixteen connections were graded *D* or *F* in Table III because the net loss was too great. Table IV shows the measured values for 1,000-cps net loss for each route, averaged for the two directions of transmission.

The over-all average of 7.5 db is reasonable, considering that the shortest connection usually had 3 switched links and that they could run to 6 switched links. The two directions differed on the average by 4 db and, in 30% of the cases, differed more than 6 db. There was a range of some 33 db from the least net loss to the greatest, and the standard deviation of net loss for the lot was 4.1 db.

AGC RANGE

All but 1 or 2 of the 16 *D* and *F* calls would have fallen into the AGC range had it been only 4 db greater or the threshold 4 db lower. Coupled with only one unsatisfactory noise rating, this suggests that a somewhat greater AGC range than 35 db (or rather a sensitivity threshold lower than −27 dbm) would be feasible and worthwhile in the Bell System plant. Since the trial, arrangements have been provided in the terminal to allow this by increasing the receiver gain on an optional basis.

TWIST

Measurements usually showed greater net loss at 2,200 cps, the carrier frequency, than at 1,000 cps. "Twist" was the name applied to this measurement of attenuation distortion, and the data system is somewhat sensitive to it. Back-to-back, the limit is about 6 db. When there is much delay distortion, twists of 2 or 3 can become important. In the field trial, the average twist turned out to be +1.5 db; the data follow a well-behaved normal law over most of the range, with a sigma of 1.5 db, but with pathological tails which made the range −3.5 to +11.5 db. High twist was blamed for 11 of the connections graded *D* or *F* in Table III. A twist equalizer (predistortion) now has been added to the transmitter. This addition has resulted in a 2- to 3-db twist correction.

LOW LOSS

As mentioned earlier, the data system returns an alarm tone when the receiving terminal fails to synchronize during a start. The discovery that there could be false alarms or even circulating alarms, therefore, was quite disconcerting. There were 14 such cases, graded *C*, which arose when there were unusual echo conditions, that is, low net losses, low return loss, and no echo suppressor. When the transmit key was released after a period of transmission, energy stored in the line could enter the terminal, which was by then ready for receiving, and thus cause false starts and alarms, to the annoyance of the user. This difficulty now has been overcome by modifying the data terminals. The plant condition is there, however, and needs to be anticipated for data systems which are not strictly one-way-at-a-time systems. For example, the problem of data correction on the alarm channel of a connection will be complicated by these low-loss situations.

Conclusions

While this trial covered a far-flung network, it represents a small sample of the connections possible in the Bell System. Nevertheless, some conclusions, however tentative, may be drawn from the experiment.

The main purpose of the trial was to demonstrate the feasibility of the switched network for the particular system described here, using a bit rate of 1,650 per second. Under trial conditions, the performance was not particularly good—only 55% of the total calls were considered good or better.

Modifications, such as elimination of false alarms because of low loss, use of predistortion, modification of compromise delay equalizer, and lowering the range of AGC when possible, were made in the terminals as a result of the trial. Therefore, an estimated 80% of these calls would be satisfactory if repeated now, assuming there is no difficulty from N overshoot. Performance can be improved still further by using longer controlled toll subscriber connections.

The results reported do not form a basis for determining general performance of the Bell System in regard to data transmission. The steps taken to overcome various difficulties for this system might be applicable to other systems. Performance on a more representative network undoubtedly would be improved on the average. Newer transmission techniques also hold promise of better service.

Worthy of emphasis is the compromise inherent in many adjustments or modifications made to improve performance. Extension of AGC range to accommodate longer (more loss) circuits increases the system's vulnerability to noise. Favoring voice-frequency links of carrier channels in designing a fixed delay equalizer is done at the expense of the amount of voice-frequency loaded circuits that can be handled. Twist equalizers can be carried only so far because of the penalty to short connections with no twist. Use of longer controlled connections between the terminal and switched network to reduce toll circuit length (and possible difficulties) in the over-all connection is probably advisable in some cases, even though the connection charge may rise appreciably.

References

1. **Capabilities of the Telephone Network for Data Transmission, A. A. Alexander, R. M. Gryb, D. W. Nast.** *Bell System Technical Journal*, New York, N. Y., May 1960.

2. **Certain Topics in Telegraph Transmission Theory, H. Nyquist.** *AIEE Transactions*, vol. 47, Apr. 1928, pp. 617–44.

3. **Assessment of Effects of Delay Distortion in Data Systems, A. D. Fowler, R. A. Gibby.** *Ibid.*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 918–23.

Modulation and Signal Selection for Digital Data Systems

ROBERT M. LERNER
MEMBER AIEE

Summary: The statistical decision model of a communication channel is used to discuss a variety of signal modulation techniques intended to overcome the problems met in the use of voice-bandwidth channels for high-performance digital data transmission. The importance of proper transmission of timing and word-start information is emphasized. The equivalence of various modulation schemes to selection from an alphabet of signal waveforms is pointed out; also that, in the presence of additive white Gaussian noise, the performance of such simple-waveform systems as AM (amplitude modulation), phase reversal, and FSK (frequency shift keying) is the same to within a few db (decibels). Implications of the fact that the channel noise is not Gaussian, but impulsive in character, are briefly explored.

IF WISHING TO CHOOSE a means of transmitting digital data from one place to another, today's communication systems engineer has available to him an ever-increasing number of devices that will accept data digits (usually binary), turn them into waveforms suitable for transmission through a given medium, deduce the transmitted symbols from the received signal, and hopefully deliver to the destination the same data symbols as were transmitted.

The engineer's selection will depend largely on his standard of performance, which, if low, will be satisfied by almost any data transmission system. When, however, the requirements are stringent, the engineer may have to answer such questions as: Can any system hope to meet the requirement? If not, why not? If so, what are the basic lines along which it must be constructed?

The purpose of this paper is to review qualitatively some ways of meeting such problems when selecting signals for use in environments similar to those encountered in the high-performance transmission of digital data through commercial and military voice-bandwidth channels. This restriction of the discussion to essentially a telephone plant environment is made, first, because the peculiarities of the voice-bandwidth channel have proved to be a limiting factor in successful operation of long-haul data transmission networks; second, because operation within the limitation of a fixed maximum bandwidth defines much of the technology that must be explored.

Nearly all audiobandwidth digital data transmission systems now being developed are intended for applications in which high performance is desired. Arbitrarily, a high-performance situation is presumed to exist whenever the communication is between semiautomatic devices in which errors are important and must be recognized or corrected automatically. The ordinary telegraph or Teletype communication of plain text messages is a good example of a low-performance system. Erroneous indication of letters in the text is common, but the human brain easily overcomes these alterations of text.

If the data to be transmitted are numerical—for instance, navigational fixes or bank balances—the detection of occasional mistakes is more difficult and, to find them, deliberate checks, such as repetition, must often be made in the transmission. So long as the checks may be reasonably performed by a human who may request of another human a retransmission to clear the garbled data, the demands upon the communication system are not stringent. However, to perform even simple checks and corrections automatically requires complicated data-processing equipment. Hence, if

---

Paper 61-1096, recommended by the AIEE Communication Theory Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted January 6, 1961; made available for printing August 14, 1961.

Robert M. Lerner is with Lincoln Laboratory, operated with support from the United States Army, Navy, and Air Force, and located at the Massachusetts Institute of Technology, Lexington, Mass.

The author wishes to acknowledge the encouragement and inspiration provided by his colleagues, Ronald Enticknap and Dr. Herbert Sherman, and express his appreciation to Valerie Sarles and Douglas St. Clair for assistance in preparing the manuscript.

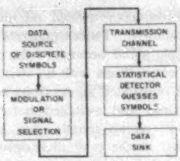

DATA SOURCE OF DISCRETE SYMBOLS
MODULATION OR SIGNAL SELECTION
TRANSMISSION CHANNEL
STATISTICAL DETECTOR GUESSES SYMBOLS
DATA SINK

Fig. 1. Model of digital data transmission system

errors must be located and discarded automatically, there is a substantial premium in building a transmission system in which the error rate is extremely low. In the automatic transmission of digital data, 1-binary-symbol error per 10^3, such symbols constitutes a high error rate; 1-bit error in 10^5 bits is considered good. Error rates in an audiobandwidth channel of less than 1 in 10^9 bits are desirable, in which case circuit interruptions for routine maintenance become more likely than errors caused by circuit noise.

Comparing the performance of various simple waveform data-signaling systems in the presence of additive Gaussian white noise results in finding that, with respect to data symbols, they are equivalent to within a few db in SNR (signal-noise ratio) required for a given error rate. A closer look at the data transmission problem reveals that (1) substantial limitations on system performance are imposed by the variable delay distortion and frequency characteristics of the chosen channel, (2) the noise is not Gaussian, and (3) proper indication of timing and word-start information is as critical as the means of indicating the data itself. Some current techniques for dealing with each of these situations are examined next.

Transmission Channel Disturbances

Freedom from errors is the most reasonable basis for evaluating the performance of a digital data transmission system. In such a case, an appropriate theoretical model of that system is the statistical decision model[1] shown in Fig. 1.

In this model, a data source generates a succession of discrete symbols selected from a finite alphabet—e.g., the upper- and lower-case symbols of a Teletype keyboard. The modulator or signal selector assigns to each symbol or group of symbols a distinct waveform suitable for transmission over the given channel. The channel effects changes in the transmitted waveform, some of them random, so that the receiver cannot always *know*, on the basis of the received signal, exactly what was transmitted, but must guess this on the same basis. This guessing process is called "statistical detection." Finally, the detected symbols, some of them erroneous, are delivered to a destination, the "data sink." Of all the assumptions that may be made about such a model, those made about the transmission channel must be the most carefully examined. What are these changes, some of them random, that the channel introduces? In the case of voice-bandwidth channels, they may be assumed to be of three kinds:

1. The channel highly attenuates all but a definite range of frequency components. Within this passband, the various frequency components of a signal may suffer unequal delays—called delay distortion or phase distortion—especially near the band edges.
2. All the frequency components of the signal may be subject to a frequency offset of heterodyning in transmission of a few cps (cycles per second).
3. The channel is disturbed by the addition of random signals to the desired one. Not only does additive noise disturb the signal, but the signal itself may fade at a rate very slow in comparison to the signal bandwidth, or suddenly drop out completely.

These assumptions do not cover all possible communication situations; indeed, they are insufficient to deal with the kind of multipath propagation and frequency selective fading typical of hf (high-frequency) ionospheric propagation. However, they are sufficient to describe the majority of data circuits in which the transmission is by wire or within-the-horizon radio relay.

Given the error rate as a standard of performance, additive noise at the receiver, and a range for the strength of the received signal, the question is: Which kind of signal selection results in the best performance? The only case where this has been answered completely is the one in which the channel disturbance is an additive noise having a Gaussian probability distribution of amplitudes. A Gaussian noise distribution is one in which the probability that, at any instant, the noise voltage lies in the range x, $x+dx$ is given by

$$(1/\sqrt{2\pi N})\exp(-x^2/2N)dx$$

where N is the average noise power.

The thermal agitation noise, quantum noise, and shot noise which limit the performance of many low-level radio receivers are of the type under discussion. For such noise, the theory[1,2] gives these results:

1. The best guessing procedure can be implemented by comparing the outputs of a set of linear filters, each matched to a symbol of the signaling alphabet.
2. If the signaling system is well designed, all binary alphabet systems (such as AM, FSK, and so forth) perform equally well, give or take a few db in equivalent signal power.

Unfortunately, the term "random noise" has become identified in many minds with the one particular kind of disturbance that has been fully analyzed. But below 100 mc (megacycles), the disturbances which limit communication system performance are terrestrial in origin and can be traced to such sources as lightning strokes, brush discharges, ignition, switching transients, and the like. Such noises are basically impulsive in character. Insofar as the probability distributions of such noises have been measured,[3–7] they have turned out to be substantially un-Gaussian, in the sense that large excursions from the rms value of the noise process are rather likely.

The error rates in systems disturbed by impulse noise cannot be deduced from a knowledge of the average SNR, measured as a ratio of signal energy to noise power per cycle, as is possible in the case of additive Gaussian noise. However, full understanding of system performance when the noise is Gaussian provides insight into what additional arrangements must be made to cope with impulsive noise disturbances.

The Signaling Process

While we have suggested a block diagram, which constitutes a *logical* description of a data transmission system, the design engineer must work from a *functional* skeleton. He needs to know what kinds of circuit functions will be essential and which incidental to the construction of any signaling apparatus.

The basic functional requirements of a data transmission signaling system are simply:

1. The physical transmission *system must have as many* recognizably different *physical states*—i.e., signaling waveforms—*as* there are *symbols in the transmission alphabet.*
2. The words "*recognizably* different" imply a *decision rule,* by which the receiver can judge the state of the system at any given time.
3. The word "transmission" implies a process occurring in time. Hence, with a *timing rule,* the receiver can know *when* the state of *the system indicates* a new *symbol.*

These basic functions must be performed by the system. Moreover, in practice, auxiliary rules and corresponding functions are often needed to take care of various housekeeping functions in the circuit, the most important of which is to set off individual words, which are groups of consecutive symbols, with a SPACE or START symbol. Such extra circuit functions add extra symbols to the symbol alphabet; in particular, the inclusion of a START symbol expands the alphabet from a minimum of 2 symbols to a minimum of 3.

The previously enumerated basic functional requirements can be illustrated with the simple telegraph system shown in Fig. 2. Here the *static* physical system has essentially two states: current flowing and current not flowing. The positive indication of the state is the closing or opening of the relay. If we now include the time dimension, the *dynamic* physical system no longer has two states. Indeed, the *change* from one static state to the other is used in conventional Morse Code to provide the *timing*; the duration of the relay closure is used to indicate which symbol, if any, is being transmitted. Thus a static 2-state system can dynamically indicate timing and at least 3 symbols—dot, dash, and space.

Questions that arise in designing this kind of data transmission systems are typical of those which beset more sophisticated apparatus:

1. How many symbols per second can be transmitted?
2. Limitations are imposed by the relay as a signal detection device, but could there be more symbols if the channels were usable as more than a 2-state device?
3. How can the duration of the symbols—that is, the timing—be specified in a sufficiently precise way?

With respect to question 1, the ultimate signaling speed of a telegraph system depends on the available channel bandwidth and the type of noise disturbances present. The corresponding theorems of Nyquist and Shannon are discussed in the next section of this paper.

Respecting use of the channel, the relay was once the only means of signal detection available, which imposed the restriction that the telegraph line could be operated as ON-OFF or by current reversal. Modern techniques have made it possible to take advantage of the continuous nature of the channel. Multiple amplitude levels,[8] multiple phase positions,[9] and even psuedo-random waveforms[8] have been used in digital data transmission.

Timing problems in data transmission are usually deceptive in their apparent simplicity; such is the case with the telegraph channel described here. Hand-sent telegraphic symbols are not distinguished by their *absolute* duration, but by their duration *relative to one another.* Moreover, in practice, these relative durations are adhered to by telegraph operators only in a nominal sense, and the difficulties in attaining automatic transcription of hand-sent Morse are discussed by Gold.[10]

If data transmission between nonhuman terminals is to be practicable, the time scale of the transmissions must be established *a priori.* At the very least, this requirement implies some sort of clock at the transmitter, an equivalent clock at the receiver, and a means of synchronizing them at least over the time duration for 1-symbol transmission of the transmitter alphabet. This is the basis of START-STOP telegraphy, or Teletype, which handles the bulk of telegraphic traffic today. Here, the initiation of the timing cycle may come at an arbitrary time by the transmission of a START pulse. The mechanisms of the transmitter and receiver are adjusted to run sufficiently well in step thereafter for the receiver to be able to identify whether the transmission is MARK or SPACE in each of 5 to 7 information time blocks, or bauds, that follow. The mechanism then, in effect, RESETS to await the transmission of the next symbol.

In the case of high-speed digital data, there is little to be gained by allowing for the asynchronous transmission of individual symbols of a message. At the very least, a transmission is made up of "words" many digits long; indeed, most transmissions consist of an esssentially endless sequence of digits. Accordingly, the "clocks" logically take on the nature of reasonably stable oscillators, which are brought into synchronization by the transmitted signal waveforms. There are, of course, cases of essentially intermittent transmission of single words, in which case the initial timing synchronization must be supplied by the equivalent of the START pulse in START-STOP telegraphy.

The rare instances in which, from an over-all system viewpoint, timing decisions ought to be made on a symbol-by-symbol basis are essential radar detection problems, which are beyond the scope of the present discussion. We reject here the example of an otherwise well-clocked transmission subject to substantial jitter—due, for example, to wow in a tape recording system. Eventually, somewhere in the data system, buffering will have to be provided so that the data may be accepted by a well-clocked device. This buffering might just as well *precede* the data transmitter and, if necessary can be built into it, thereby simplifying the design of the transmission equipment.

Simple Modulation Systems

A simple modulation system is one in which the channel signal is varied between states—usually two, called MARK and SPACE—in some convenient region of the spectrum other than direct current or low-pass. The systems now commercially available vary in sophistication all the way from those that do little more than accept and deliver a 2-state d-c signal to those which accept and deliver data symbols, timing information, and a word-start or segmentation symbol. Typically, some form of amplitude, phase, or frequency modulation is used to translate the spectrum of the input data signal into a usable portion of a voice channel, or other transmission facility, passband. Rather than discuss each possible system in detail, several general questions are investigated:

1. Limitations on the rate of signaling that may be achieved over a given channel.
2. Practical techniques for exploiting available channel bandwidth.
3. The equivalence of the modulation and signal selection point of view.
4. Potential equivalence of all simple modulation signals in the presence of additive white Gaussian noise to within a few db with phase modulation holding a slight edge
5. The means used—or which should be used—to recover timing information.
6. Provisions that are—or can be—made for the transmission of the word-start symbol.

RATE OF SIGNALING

How fast can data be transmitted over a channel of a given bandwidth? The answer depends on one's thinking. If it is in terms of easy instrumentation, the answer is provided by Nyquist's theorem, the sampling theorem, which roughly states that in a channel whose positive-frequency bandwidth is limited to W cps, it is possible to transmit $2W$ independent samples per second. If wanting to know the possible ultimate transmission rate, one needs to learn something about the statistics of the channel noise. Roughly speaking, the noise may be re-

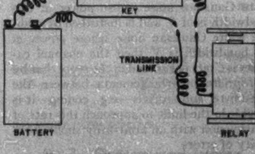


Fig. 2. Elementary telegraph

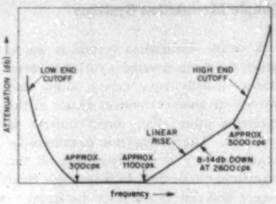


Fig. 3. Relative attenuation characteristic of switched plant telephone circuits; after Alexander, Gryb, and Nast[13]

garded as quantizing each independent sample amplitude into levels separated by the rms amplitude of the noise. The number of binary digits or bits, equivalent to each such sequence of levels, is

$$\log_2 \sqrt{1+S/N} \text{ bits per sample} \qquad (1)$$

where S/N is the signal-to-noise ratio in power. Thus, it is possible to transmit roughly

$$C = 2W \log_2 \sqrt{1+S/N} = W \log_2 (1+S/N) \text{ bits per second} \qquad (2)$$

or binary digits per second over a channel whose bandwidth is W.

Some precautions are necessary in the interpretation of Nyquist's theorem and equation 2, Shannon's theorem.[11]

1. Both theorems are, strictly speaking, based on the assumption of a channel whose response in the passband is flat and whose attenuation outside the passband is infinite. From a practical point of view, failure to meet these conditions precisely is not serious, since the channel response near a nominal band edge is likely to be unpredictable, particularly in multilink transmission systems. Further, a channel will generally exhibit severe delay distortion near the band edges. The variability and difficulties in fully compensating these departures from ideal passband characteristics restrict the usable bandwidth of transmission channels to a fraction of the nominal bandwidth—usually about 1/2 bandwidth in the case of multilink voice-bandwidth channels.

2. The derivation of equation 2 suggests a high probability of confusion between the various quantized sample levels in that the sample-step SNR is unity; this is *not*, however, the information theory interpretation of equation 2. Shannon showed that if a continuous channel has an ideal bandwidth of W cps, is limited to an average signal power of S, and is disturbed by additive white Gaussian noise whose power in this bandwidth is of S, and is disturbed by additive white Gaussian noise whose power in this bandwidth is N, then the channel capacity is C bits per second in the sense that by appropriate prearrangements between the transmitter and receiver—e.g., coding—it is possible in the limit to approach this rate of transmission with an arbitrarily small probability of errors.

Thus far in the state of the art, such coding arrangements have been quite complicated where they are most useful—namely, when the error rate becomes momentarily high. However, considering the complexity of the terminal equipment often served by data transmission systems, the engineer should not lightly discard the possibilities of error-correction coding. Recent developments in coding theory suggest that coding for error *detection*, rather than *correction*, is both practicable and useful.

In view of the sophistication of the coding arrangements required to realize error-free communication at close to the Shannon rate, equation 2, efforts to exploit the full bandwidth of a channel have generally been in the direction of a closer approach to the transmission of $2W$ independent samples per second.

This latter problem has been attacked both by compensating known channel spectrum distortions—principally time-delay distortion, and by shaping the spectrum of the modulation output signal—principally the vestigial sideband arrangements. More and more data are becoming available with respect to the transmission characteristics of the voice telephone plant. There exist so-called "data grade" circuits which are compensated to meet specifications[12] designed to insure the availability of 1.5 kc of bandwidth, for which the variations in channel attenuation and delay are less than ±3 db and 50 μsec (microseconds), respectively. Alexander, Gryb, and Nast[13] have recently reported a series of measurements on urban and interurban switched plants. Some of their results, relating to typical attenuation and delay distortion, respectively, are given in Figs. 3 and 4.

These data suggest that, on the average, usable bandwidth might be gained by providing delay and amplitude compensation in the data receiver. Both fixed and adjustable compensation have been built into equipment now, or soon to be, commercially available—the former correcting the distortion of the average circuit and the latter set to meet the requirements of a specific one.

No good general answer has been forthcoming as to how serious is the effect of given amounts of amplitude or delay distortion. With respect to delay distortion, some workers[14,15] have suggested a criterion involving the jitter introduced into the signal appearing at the output of an approximate envelope detector. A good general rule-of-thumb is that the variation in delay ΔT across the passband for a signal W-cps wide ought not to exceed $1/W$ to $2/W$ seconds.

Two further techniques have been employed to make the most of the available bandwidth. One seeks to eliminate the effect of delay distortion by dividing the channel bandwidth W reasonably available—from an attenuation point of view—into a set of N narrow-band subchannels, taking N sufficiently large so that

$$\Delta TW/N << 1 \qquad (3)$$

The data are then transmitted in groups of N symbols in parallel over the narrow band subchannels. The subchannels exhibit differences in transmission delay in a range of ΔT seconds, but, on account of the narrow subchannel bandwidth seen in equation 3, this variation is inconsequential compared with the duration of a symbol waveform. Such variations are easily removed by subsequent decision circuits.

The subchannel method of combating phase distortion requires, in principle at least, the availability of as much or more filtering than is needed to compensate the amplitude and delay distortion of the channel; this was suggested earlier. Unless the data source naturally produces data in N parallel channels, the use of the subchannel system requires a pair of serial-to-parallel conversion operations at the terminals. From the viewpoint of immunity to channel phase distortion alone, the subchannel system has no inherent superiority to direct compensation of the channel at one or both bands.

To permit efficient utilization of the channel bandwidth, the common practice is to eliminate redundant regions of the transmitted signal spectrum by vestigial and single-sideband techniques. All simple modulation schemes produce a signal which has two sets of sidebands, one above and one below the carrier frequency. Only one set is necessary to convey signal information. Considering the spectrum of the simple ON-OFF signal shown in Fig. 5, and assuming that it is ON as often as it is OFF, half the power of this signal is in a d-c line at zero-frequency. The remaining power is divided equally between a spectrum from zero to

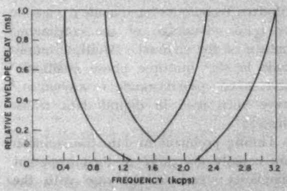


Fig. 4. Switched plant envelope delay distortion characteristics—locus of 90% of all circuits tested; after Alexander, Gryb, and Nast[13]

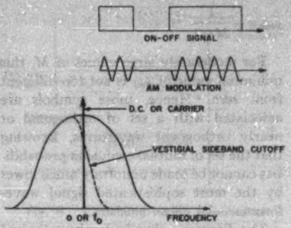


Fig. 5. Spectrum of modulated signal

W cps and a minor image of this spectrum from zero to $-W$ cps. If this ON-OFF signal is used to modulate a carrier at f_0 cps, then the *positive* frequency part of the modulated waveform spectrum is exactly the same as that shown in Fig. 5, except that zero-frequency has now become f_0.

From an information standpoint, the d-c or carrier term is unessential, represents wasted power, and can be eliminated. At carrier frequency, this results in suppressed carrier modulation, or, as it is better known in data transmission, a phase reversal system. The 3 db of power that was formerly in the carrier may now be used in the information-carrying sidebands; hence, for a given signal power, a phase reversal system is in principle potentially 3 db better than an AM system. Whether or not the carrier is eliminated, it is customary to save spectrum by transmitting only one of the information-bearing sidebands. Since there is usually no need to completely eliminate all of the superfluous spectrum, the transmitted spectrum resembles that of a vestigial rather than a strict single sideband signal, as suggested by the dashed line in Fig. 5.

The FSK is basically double amplitude modulation, using two different carriers, one being keyed ON while the other is OFF. The spectra about these two carriers overlap by an amount which depends on the carrier separation; see Fig. 6. Clearly, the bandwidth required for this kind of transmission is greater than that needed for AM by an amount at least equal to the carrier separation. The advantage over AM is that there is no necessity to keep track of phase relations or amplitude levels at the receiver; the amplitude of the f_1 signal acts as the reference level for the other, and vice versa. As before, the superfluous sidebands—in this case the outside pair—need not be transmitted. Also as before, a 3-db improvement in transmitted power may be had by eliminating the carriers, although the resulting double phase reversal system had never been tried to the author's knowledge.

Equivalence of Modulation and Signal Selection

In all of the modulation schemes just described, it is possible to recover the original ON-OFF waveform at the receiver if the signal has not suffered too much from adverse transmission conditions. However, the problem at the receiver is to recognize which symbol was transmitted. For this purpose, it is not necessary to recover the premodulation form of the signal; rather, it suffices to know what waveform is expected at the receiver for each of the possible transmitted symbols. This fact suggests that we view the original signal production, not as a modulation process but rather as one selection from an alphabet of signal waveforms, each corresponding to a data symbol.

If, for instance, each of the carrier frequencies, f_1 and f_2, is an integral multiple of the digit rate, then the waveform corresponding to MARK or SPACE always starts with the same initial phase. We might just as well have obtained it by impulsing a linear network whose transient response is a rectangular pulse of sine wave at frequency f_1 each time a MARK comes along and a similar pulse at frequency f_2 for a SPACE.

Thus, in general, we replace the idea of signal modulation by the idea of assigning to each of N possible different symbols a unique waveform, $s_1(t)$, $s_2(t)$, ... $s_n(t)$, which may be as simple or as complex as we choose. Of course, if in the example of FSK there is no harmonic relationship between the carriers and/or the digit rate, there is still a question of the initial phase of the waveforms $\{s_k(t)\}$ to be settled. However, such questions are not critical. Indeed, in providing for the presence of 1-cps to 10-cps phase crawl (more in older installations) in the receiver symbol detection circuits, variable initial phase angle in the $\{s_k(t)\}$, due to the use of modulators, is automatically provided for.

Detection of Data Disturbed by Gaussian Noise

The optimum procedures for detecting—i.e., guessing—signals in the presence of additive white Gaussian noise are well understood. Here, the results of decision theory guide the design of signals that are to be used in channels in which additive white Gaussian noise is the primary disturbance. Only a few special cases of signal selection for this kind of noise environment are discussed.

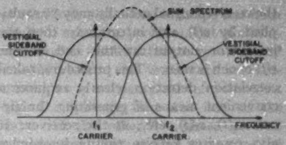


Fig. 6. Spectrum of FSK signal

In the simplest case, the choice is to be made between two symbols uniquely specified at the receiver by $s_1(t)$ and $s_2(t)$. Assuming that the symbols are equally likely; that the cost of an incorrect guess is independent of the symbol transmitted; and that the waveforms $s_1(t)$ and $s_2(t)$ have equal energy E, then

$$\int s_1^2(t)dt = \int s_2^2(t)dt = E \tag{4}$$

Under these circumstances the probability that $s_1(t)$ was transmitted and given the reception of $y(t)$ is essentially the probability that the added noise, $n(t)$, was

$$n(t) = y(t) - s_1(t) \tag{5}$$

If $n(t)$ is Gaussian, its probability is inversely dependent on its energy, that is

$$\text{prob}\,(n(t)) \text{ varies as } -\int n^2(t)dt \tag{6}$$

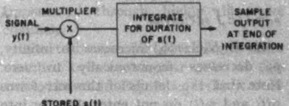
The decision between $s_1(t)$ and $s_2(t)$ as the original signal will then depend on whether the noise $y(t) - s_1(t)$ or the noise $y(t) - s_2(t)$ is the more probable; on whether the left or right hand sides of equation 7 is greater.

$$-\int (y(t) - s_1(t))^2 dt \gtrless \int (y(t) - s_2(t))^2 dt \tag{7}$$

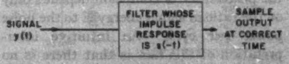
Squaring out the integrands in equation 7 and making use of equation 4, the question is whether the left or right hand side of equation 8 is greater.

$$\int y(t)s_1(t)dt \gtrless \int y(t)s_2(t)dt \tag{8}$$

The integrals in equation 8 may be evaluated in one of two ways. First, as illustrated in the upper part of Fig. 7,


MULTIPLIER
SIGNAL
y(t)
X
INTEGRATE FOR DURATION OF s(t)
SAMPLE OUTPUT AT END OF INTEGRATION
STORED s(t)

(a) CORRELATION DETECTION


SIGNAL
y(t)
FILTER WHOSE IMPULSE RESPONSE IS s(−t)
SAMPLE OUTPUT AT CORRECT TIME

(b) MATCHED FILTER DETECTION

Fig. 7. Optimum signal detectors

the received signal actually may be multiplied by $s(t)$ and integrated; then the integrator output at the end of the $s(t)$ epoch is seen. This procedure, called correlation detection, clearly requires a convenient means of generating the signals $s_1(t)$, $s_2(t)$, etc., at the receiver; it also requires precise timing knowledge concerning the received symbol waveforms.

The second method takes advantage of the convolution performed by a linear filter. If a signal $y(t)$ is passed through such a filter whose impulse response is $g(t)$, the filter output $z(t)$ is given by the integral

$$z(t)=\int_{-\infty}^{\infty} y(x)g(t-x)dx \tag{9}$$

Accordingly, by passing $y(t)$ through a filter whose impulse response is $s(-t)$, the filter computes the integral

$$z(t)=\int_{-\infty}^{\infty} y(x)s(x-t)dx \tag{10}$$

At $t=0$, this is precisely the integral equation 8 needed in deciding which symbol was most probably transmitted. This method of signal processing, illustrated in the lower part of Fig. 7, is called matched filter detection. When using this method, it is not necessary to know in advance exactly when the signal occurs; the filter computes the required integral continuously in time. Opposing this advantage is the disadvantage that the matched filters may be difficult to construct if the signals $s_1(t)$, $s_2(t)$ ... are complicated.

With either method of deciding which symbol was transmitted, it is possible to compute the probability of confusing one symbol with another. When the Gaussian noise energy is N_0 watts per cycle, the probability of such error is

$$p_c=\text{prob of confusion}=\Phi[(E/N_0)(1-\rho_{12})]^{1/2} \tag{11}$$

in which Φ is the normal error integral

$$\Phi(x)=1/\sqrt{2\pi}\int_x^{\infty} \exp(-\zeta^2/2)d\zeta \tag{12}$$

and

$$\rho_{12}=1/E\int s_1(t)s_2(t)dt \tag{13}$$

As $E/N_0(1-\rho_{12})$ increases to infinity, p_c decreases monotonically to zero. Note that the details of the waveforms $s_1(t)$ and $s_2(t)$ do not enter directly into equation 11 to cause probable confusion. Neglecting momentarily the factor $(1-\rho_{12})$, which can vary only between 0 and 2, the probability of confusion depends only on the ratio of signal energy E to the noise power per cycle. An intuitive interpretation of this fact is that there is no signal waveform that looks least like Gaussian noise; for any value of ρ_{12}, Gaussian noise will turn any one waveform into any other waveform of the same energy with equal probability.

If transmitting signals at a fixed rate of n per second, and with a fixed amount of available average power, the energy per symbol E is fixed also. In this case, the only remaining design parameter is ρ_{12}, which clearly should be as negative as possible so that the argument of the error function in equation 11 can be as large as possible. The choice of $s_1(t)$ and $s_2(t)$ is made accordingly.

The results obtained up to this point can help to answer an often-asked question, "Which is the best system for digital-data transmission: frequency shift keying, phase reversal, or AM?" Actually, none of these is best in the sense of being an ultimate data communication system, but all can be compared with respect to the Gaussian noise environment. The use of FSK at a given power level corresponds to choosing ρ_{12} equal to zero, since the average product of two functions is zero if their spectra occupy different frequency bands. In this circumstance, the probability of confusion is

$$p_c=\Phi(E/N_0)^{1/2}=\Phi(S/nN_0)^{1/2} \tag{14}$$

where S is the average power of the transmission. But the orthogonal-signal value $\rho_{12}=0$ is not the smallest value for ρ_{12} design. Taking $s_2(t)=s_1(t)$ gives $\rho_{12}=-1$, in which case the system is shown to be 3 db better off than it was in the case of FSK or other orthogonal signals

$$p_c=\Phi(2E/N_0)^{1/2}=\Phi(2S/nN_0)^{1/2}$$

This 3-db improvement over orthogonal signals, it should be emphasized, is a potential property of *any* waveform reversal system, of which the carrier phase reversal system is but a special case. Further, this result holds true only for additive Gaussian noise channel disturbances. As pointed out before, an AM system is also at least 3 db inferior to a waveform reversal system.

The decision so far has been between only two alternative symbols. Now, suppose there are M symbols whose corresponding waveforms are $s_1(t)$, $s_2(t)$... $s_M(t)$ which have equal energy. Unfortunately, for $M>2$, no simple formula for the probability of confusion exists and, for $M>3$, integrals appear which have not been tabulated. Accordingly, the probabilities of pair-wise confusions are made as low as possible. By considering the properties of a regular polyhedron having M vertexes in $M-1$ space, it can be shown^2 that the most negative value that can be simultaneously taken by the correlation coefficients between members of an M waveform alphabet is

$$\rho_{mk}\triangleq-\frac{1}{M-1}$$

For moderately large values of M, this minimum value of ρ_{mk} is not too different from zero. Hence, those symbols are associated with a set of orthogonal or nearly orthogonal waveforms, knowing that the set of mutual confusion probabilities cannot be made uniformly much lower by the most sophisticated signal waveforms.

The discussion thus far has been limited to a decision made at one fixed time, between symbols that were possibly transmitted at some other fixed time. Since in practice a succession of symbols is transmitted, it is necessary to make a succession of decisions at the receiver. In this connection, details about the waveform of $s_1(t)\ldots s_M(t)$ are important—particularly, details about the effective bandwidth. In principle, the waveforms due to all likely sequences of symbols should be treated as symbols in making a statistical decision about the transmission. The difficulty with doing so is that the amount of equipment needed at the receiver grows exponentially with the number of symbols in a sequence.

To avoid the need for this kind of complexity in the receiver, the signal produced in the output of any matched filter by the transmission of a symbol must be zero or nearly zero whenever decisions are made, excepting for one time when transmission of that particular symbol is at issue. Unless this is done, the decision at any one time will be interfered with by the effects of tails of signals transmitted at other times, and the system tends to jam itself. One way to prevent this intersymbol crosstalk is to choose waveforms $s_1(t)$, $s_2(t)\ldots s_M(t)$ whose duration is sufficiently short. The general requirement is not that the signal be short, but that its bandwidth be large enough that the correlation functions can be specified approximately at intervals equal to the time between signals. In choosing signals for a channel of strictly limited bandwidth, it is impossible to avoid signals which overlap in time.

Matched filter detection has been considered, but without including such practical questions as physical realizability. Clearly, a filter cannot be built whose impulse response is $s(-t)$ if the implication is that the filter exhibits response before it is excited. Actually, any filter will perform nearly as well as a matched filter—usually to within a fraction of a db in equivalent signal power—if the filter passband has approximately the same shape as that of the signal spectrum, and if the processed signal emerging from the filter has a linear

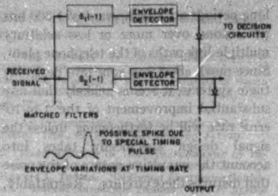


Fig. 8. Obtaining timing information

phase to within ±90 degrees over the effective signal bandwidth. Of these two requirements, the latter is more stringent; it amounts to introducing sufficient delay into the filtering to make this operation physically realizable.

The system error rate is relatively insensitive to moderate differences between the *magnitude* of the frequency response of an actual receiver filter and that of the ideal matched filter. Accordingly, the receiver filters may be deliberately somewhat mismatched in order to permit the use of less expensive circuit components or to insure that the filter output goes to zero whenever the next symbol is to be indicated.

TIMING

Both correlation detection and matched filter methods require precise knowledge as to when to compare the outputs of various signal processing filters. The ways of indicating timing to the data receiver are to: (1) transmit timing information on a carrier or subcarrier along with the data, (2) derive the timing by extrapolation from the occurrence of a special START symbol, and (3) derive timing from the envelopes of outputs of the signal processing filters.

If a transmission system is subject to phase crawl, a timing carrier waveform is difficult to use therein. Insufficiently well-locked oscillators in single-sideband links of the transmission path effectively change the frequency of the timing carrier without changing the data rate. Thus, a minimum of two frequency components is needed in the timing waveform. Even then, the maintenance of proper phase relationships between these components is not insured through the phase distortion of the transmission path. Thus, the best set of frequency components to indicate the timing is the data signal itself, as suggested in alternative 3.

The extrapolation of timing from the occurrence of a START symbol is commonly used when the number of intervals to be timed is not too large. In START-STOP telegraphy, the timing is basically mechanical. In at least one digital data system that has been used,[18] the timing was provided by impulsing a high-Q resonator whose subsequent ringing supplied the timing. To obtain timing from occasional START symbols places severe constraints on the stability of the elements in the data transmission sets. When R—the number of symbols to be timed per START symbol—exceeds a few dozen, the devices determining the time scale at transmitter and receiver must maintain their relative stability to within a fraction of $1/R$.

Only in the special case where equal amplitude symbols are being transmitted through a strictly bandlimited channel at a rate equal to twice the channel bandwidth is it theoretically impossible to deduce timing information from the data signal itself. Otherwise, in principle, an optimum estimate of the timing can be made from an examination of the received signal. Rather than carry through a detailed theoretical investigation, note that the filter matched to each successive indicated symbol has a maximum output at the time the decision is made. These peaks occur, with jitter, at roughly the correct timing point. Hence, the continual examination of the largest envelope output of all the matched filters produces a quasi-periodic waveform, whose peaks on the average specify the correct timing. See Fig. 8.

The best way to average out this rough timing data is to use the information in synchronizing and phase-locking a timing oscillator in the data receiver. With such an arrangement, the data are averaged over many symbols without prejudicing the ability of the timing device to adapt to a reasonable range of data rates. Such averaging can also be performed by fixed tuned circuits, but such a procedure suffers from the difficulty referred to earlier, namely, that a high degree of stability is required in the equality of the transmitted data rate and the resonator center frequency.

Fig. 8 shows the possible presence of a special occasional timing signal, such as the START symbol discussed previously. Such occasional timing waveforms can be used to phase-lock a local timing oscillator, although noise-induced jitter in such occasional signals will critically affect the proper operation of the timing circuits. Accordingly, it is usually better to work with the maximum envelope of the matched filter outputs.

START SYMBOL

Relatively little attention has been given to proper transmission of the START symbol in the design of digital data systems.[17] In at least two systems presently in use, some attempt has been made to meet the problem, but the solutions are not wholly satisfactory. In one,[3] an AM system, the word-start is transmitted as a carrier pulse of somewhat higher amplitude than the rest of the data. In the other,[18] a phase reversal system, the START symbol is transmitted as a unique sequence of 5 data symbols and serves as the word-start. In order that this sequence be unique, the system contains a code translation stage which converts every group of 4 input data symbols to 5 for transmission; a 5-to-4 retranslation is performed in the receiver.

Both systems are so conceived as to give the word-start more protection from error than is accorded to a symbol within the word, but they do not go far enough in this direction. The AM system is particularly susceptible to false indication of word-start implemented by impulse noise.[3] Much more work needs to be done in devising means of transmitting well-protected word-start pulses.

Other Signaling Techniques Based on Gaussian Theory

When the channel noise is Gaussian—which usually means higher than 100 mc—there are only two basic ways of improving a decision's reliability: by building more perfectly matched filters and by increasing the energy per decision through a multisymbol alphabet of transmitted waveforms. Actually, neither technique buys much beyond a certain rapidly reached point, because, in the respective cases, the detection procedure is insensitive to moderate mismatch in the receiver filters, and the signaling alphabet size tends to grow exponentially with the increase in integrated energy. Nevertheless, a brief examination is here accorded these techniques since they are indicative of the technology now developing for dealing with signals in the presence of impulse noise[3] and signals that must be transmitted through highly dispersive channels.

One technique for obtaining a filter matched to a rectangular pulse of sine wave, shown in Fig. 9, is really a com-

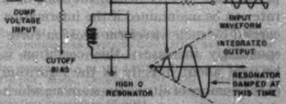


Fig. 9. Integrate and dump circuit

bination of matched filter and correlation detection ideas since the signal processing is done with a time-varying circuit, in this case a high-Q resonator whose Q may be suddenly lowered by the application of a control voltage—in this case, bringing a cathode follower from the cutoff to the conducting state. The level of oscillation in the unclamped resonator builds up linearly with time as the appropriate rectangular envelope sinusoid arrives. At the end of the pulse of sinusoid, the resonator continues to oscillate at the end-point level until it is damped, or dumped. As a matched filter alone, this device is probably overdesigned. A similar resonator with the correct moderate Q as a mismatched receiver filter would do as well to within 1 db; but it conveniently discriminates against other orthogonal waveforms in a multisignal alphabet of pulsed carrier waveforms.

A high signal-to-noise ratio may evoke a willingness to trade part of the SNR for an increase in data rate. To engineer this trade, simply construct a multisymbol alphabet by quantizing the amplitude or the initial phase of a given symbol waveform into a number of recognizably different steps. This procedure is prodigal in its use of available SNR, for the equivalent binary symbol data rate increases only logarithmically with the number of quantum steps. Nevertheless, this quantizing technique has been widely used to increase the data rate by a factor of 2 or 3 beyond the transmission rate of independent waveforms. Usually, phase quantization is chosen rather than amplitude quantization because of the variability of the net loss characteristic of the transmission channels.

Quite different is the use of a multisymbol alphabet for *increasing the reliability* of the data transmission.[19] Suppose that the average power and number of symbols to be delivered per second to the receiver are given, and that ρ_{12} (see previous main section of this paper) has been properly selected. With all these selections, the error rate is still too high. What can be done to improve it further? Clearly, rise in the energy integrated for each decision is dictated.

More energy per decision means, in effect, the transmission of fewer symbols per second. Nevertheless, the same data rate can be maintained if the information carried by each waveform decision correspondingly increases; this condition is brought about directly as the logarithm of the number of alternatives among which the choice is made. Therefore, a larger signal alphabet will be able to transmit the same information with fewer symbols per second. This increase in alphabet size must be of a different kind than that mentioned in the last paragraph; for the symbol waveforms must remain nearly orthogonal (again, see previous main section of this paper) if the probability of confusing them is to decrease as the energy is raised. If orthogonality is maintained as the energy integrated per decision is increased indefinitely, the probability of error at any data rate up to Shannon's channel capacity decreases to zero.[20] The signal waveforms of the exponentially large alphabet required to approach this error-free condition are in the limit complicated, perhaps resembling samples of Gaussian noise whose duration bandwidth product greatly exceeds 1.

If the symbol waveforms are orthogonal (all $\rho_{mk}=0$), all intersymbol confusions are equally likely. If p_c, the probability of a pair-wise confusion, is small to start with, then the probability of a confusion in the case of an M symbol alphabet P_m is given approximately by

$$p_M \doteq (M-1)p_c \tag{15}$$

In particular, in the case of a decision between M orthogonal symbols, the probability of confusion is

$$p_M \doteq (M-1)\Phi\left(\frac{E}{N_0}\log M\right)^{1/2}$$

for values of the argument greater than about 5 db. Thus, the effectiveness of using a large number of symbols in the alphabet grows at best logarithmically with the size of the alphabet. In going from a binary alphabet to one with 32 symbols, an approximate 6-db gain (3 db with respect to phase reversal) is realized if E/N_0 is greater than about +5 db. Since the gain in effective signal energy increases so slowly with the number of symbols in the alphabet, the use of alphabets whose size ranges from a dozen to 100-odd symbols seems to be the practical limit of exploiting this particular avenue to increasing the reliability of the channel without affecting the rate at which data are delivered. In case further reductions in error rate are necessary, a digital coding scheme probably offers the best promise for success.

Future Developments: Impulse Noise

The most diligent combinations of the best of the modulation and signaling techniques which have been described in part herein have yet to produce a data system which averages much fewer than 1 binary error in 10^5 bits when operated for long periods at rates upward of 1,300 bits per second over more or less arbitrary multiple-link paths of the telephone plant. Since much effort has gone into designing these systems, it seems unlikely that any substantial improvement of the 1-in-10^5 error rate will be forthcoming unless the signal design specifically takes into account the impulsive nature of the noise that disturbs these circuits. Regrettably, not enough is yet known of the statistics of such noise to rigorously guide the designer of signal selection and detection devices. Knowledge of impulse noise is rapidly increasing as can be seen by consulting reference 3.

Despite inadequate knowledge about details of impulsive noise distributions, they may be expected to exhibit two characteristics which are exploitable in signal design and detection:

1. The mechanism by which such noise is produced usually can be traced back through a simple physical system to a source which, for all practical purposes, produces true mathematical impulses. Accordingly, the observed noise waveform resembles the transient response of a simple physical system; namely, that between the pulse source and the observer.
2. The noise source delivers energy in bursts rather than smoothly in time.

With respect to the first characteristic, transmission protection is sought by choosing signal waveforms that bear little resemblance to the expected noise waveforms; in this connection, it may even prove possible to identify the noise whenever it occurs and substantially to eliminate its effect.

As to the second characteristic, in systems generally dealt with, the *average* noise power which might be delivered in the output of a matched filter is so low that the noise energy appearing at the point where decisions must be made is delivered at a low average rate instead of in bursts.

The two types of waveform design just suggested are supplementary so that, under some circumstances, a combination of the two can virtually eliminate the effect of impulse noise. Encouraging laboratory work[7] is in progress along both the lines suggested; hopefully, modulation and signaling techniques will soon be able to protect the transmission of audio-bandwidth digital data from all but the most catastrophic standard telephone channel disturbances, such as circuit interruption.

References

This listing is not intended to be an exhaustive survey of the literature available on digital data modulation and signaling systems.

1. The Theory of Signal Detectability, **Birdsall, Peterson.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, New York, N. Y., vol. IT-4, Sept. 1954, pp. 171–212.

2. An Introduction to Matched Filters, **G. L. Turin.** *Ibid.*, vol. IT-6, June 1960, pp. 311–29.

3. Some Results in the Measurement of Impulse Noise on Several Telephone Circuits, **Howard L. Yudkin.** "National Electronics Conference, Data Transmission Session," *AIEE Special Publication T-123*, 1960, pp. 65–74.

4. Measured Error Distributions on the Bell A-1 Facility Over Various Media, **Erhard J. Hofmann.** *Ibid.*, pp. 57–64.

5. Data Transmission Performance on Long-Haul Telephone Facilities, **Paul V. Dimock.** *Ibid.*, pp. 23–30.

6. Model of Error Burst Structure in Data Transmission, **Pierre Mertz.** *Ibid.*, pp. 75–83.

7. Lectures on Communication System Theory (book), **E. Baghdady, editor.** McGraw-Hill Book Company, Inc., New York, N. Y., 1961, especially chap. 11.

8. SAGE Data Terminals, **R. O. Soffel, E. G. Spack.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 872–79.

9. Kineplex, a Bandwidth-Efficient Binary Transmission System, **R. R. Mosier, R. G. Clabaugh.** *Ibid.*, vol. 76, 1957 (Jan. 1958 section), pp. 723–28.

10. **Gold.** *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, vol. CT-5, no. 1, Mar. 1959, pp. 17–24.

11. Communication in the Presence of Noise, **C. E. Shannon.** *Proceedings*, Institute of Radio Engineers, vol. 37, Jan. 1949, pp. 60–93.

12. SAGE Data System Considerations, **R. G. Enticknap, E. F. Schuster.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 824–32.

13. Capabilities of the Telephone Network for Data Transmission, **A. A. Alexander, R. M. Gryb, D. W. Nast.** *Bell System Technical Journal*, New York, N. Y., vol. 39, May 1950, pp. 431–76.

14. A 2500 Baud Time Sequential Transmission System for Voice Frequency Wire Line Transmission, **G. Holland, J. C. Myrick.** *National Convention Record*, Institute of Radio Engineers, pt. 8, 1959, pp. 187–96; also *Transactions*, Professional Group on Communications Systems, Institute of Radio Engineers, vol. CS-7, Sept. 1959, pp. 180–84.

15. An Evaluation of AM Data System Performance by Computer Simulation, **R. A. Gibby.** *Bell System Technical Journal*, vol. 39, May 1960, pp. 675–704.

16. Synchronized Clocks for Data Transmission, **J. O. Edson, M. A. Flavin, A. D. Perry.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 832–36.

17. The Design of Digital Data Transmission Systems (book), **R. G. Enticknap, R. M. Lerner.**

18. A New Phone Line Data System for Use in SAGE, **K. E. Perry.** *Group Report 32–105*, Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.; also *ASTIA no. 230115*, Armed Services Technical Information Agency, Dayton, Ohio.

19. The Asymptotic Behavior of Ideal M-ary Systems, **G. L. Turin.** *Proceedings*, Institute of Radio Engineers, vol. 47, Jan. 1959, pp. 93–94.

20. Probability of Error for Optimal Codes in a Gaussian Channel, **C. E. Shannon.** *Bell System Technical Journal*, vol. 38, May 1959, pp. 611–56.

21. Some Results on the Transmission of Pulses Over Telephone Lines, **J. V. Harrington, P. Rosen, D. A. Spaeth.** *Symposium on Information Networks-III*, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., Apr. 1954, pp. 115–130.

22. A Frequency-Modulation Digital Subset for Data Transmission Over Telephone Lines, **Laurance A. Weber.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 867–72.

23. Binary Data Transmission Techniques for Linear Systems, **M. L. Doelz, E. T. Heald, D. L. Martin.** *Proceedings*, Institute of Radio Engineers, vol. 45, May 1957, pp. 656–61.

24. An Experimental Modulation-Demodulation Scheme for High-Speed Data Transmission, **E. Hopner.** *IBM Journal of Research and Development*, New York, N. Y., vol. 3, Jan. 1959, pp. 74–84.

25. Transmission of Digital Information Over Telephone Circuits, **A. W. Horton, Jr., H. E. Vaughan.** *Bell System Technical Journal*, vol. 34, May 1955, pp. 511–28.

26. Transmission Aspects of Data Transmission Service Using Private Line Voice Telephone Channels, **P. Mertz, D. Mitchell.** *Ibid.*, vol. 36, Nov. 1957, pp. 1451–86.

27. Assessment of the Effects of Delay Distortion in Data Systems, **A. D. Fowler, R. A. Gibby.** *AIEE Transactions*, pt. I (*Communication and Electronics*), vol. 77, 1958 (Jan. 1959 section), pp. 918–23.

28. Modulation Theory (book), **Harold S. Black.** D. Van Nostrand Company, Inc., Princeton, N. J., 1953.

29. The Mathematical Theory of Communication (book), **C. E. Shannon.** University of Illinois Press, Urbana, Ill., 1949.

30. Double Sideband vs. Single Sideband Systems, **J. Costas.** *Proceedings*, Institute of Radio Engineers, vol. 45, Apr. 1957, pp. 534–37.

31. Integrated Data Systems, **R. Filipowsky.** *Transactions*, Professional Group on Communications Systems, Institute of Radio Engineers, vol. CS-7, June 1959, pp. 95–101.

32. Impulsing Linear Networks in Integrated Data Systems, **G. K. McAuliffe.** *Ibid.*, Sept. 1959, pp. 189–95.

33. Performance of Digital Phase Modulation Communications Systems, **C. R. Cahn.** *Ibid.*, May 1959, pp. 3–6.

Digital Tropospheric Scatter

J. R. POPPE
NONMEMBER AIEE

A. J. VADASZ
NONMEMBER AIEE

Summary: An old communications tool is put to new use by combining it with new concepts in propagation. Using simple, proven binary detection methods—such as frequency shift keying—substantial range improvement is shown over multichannel FM (frequency-modulated) or SSB (single-side-band) tropospheric scatter schemes, at the sacrifice of analog capability. Alternatively, over shorter distances, substantial reduction in equipment complexity can be achieved over other modes. Measurements made on a feasibility circuit of 150 miles using very simple equipment has shown the practicality of this concept.

Trends in Tropospheric Scatter

THE TRENDS in tropospheric scatter communication systems have been toward higher channel capacities and greater hop lengths. One installation has a capacity of 132 voice channels.[1] Another has a hop length of 690 miles with a capacity of 24 voice channels. Other systems in the planning stage have capacities of 240 voice channels over 150-mile distances. Extremes in these parameters are noncurrent however. Terminal and repeater stations of substantial size are required as well as large antennas, high-power transmitter amplifiers, and high orders of diversity. These trends have evolved due to pressing communications' needs and have brought about the employment of novel devices and systems techniques.

It has become apparent, however, that beyond the radio horizon UHF (ultra-high-frequency) tropo-scatter relays are still system margin limited. In other words, power output devices and highly directional antennas are the cost governing factors. Therefore, it is reasonable to expect that an approach in the opposite direction, such as reduction in the size, cost, and complexity of these components, is warranted even at the expense of restricting the type or capacity of information handling potential. Investigations in this area have produced interesting and useful results.

Theoretical Considerations of Digital Tropospheric Scatter

Our starting point has been binary detection, using both digital data and coded information as the base signal as in teletypewriter information. The potential rewards of such an arrangement are illustrated in Fig. 1. Curve *B* is the theoretical error rate performance of a frequency shift keyed (FSK) binary channel using an optimized detector as a function of carrier-to-noise ratio (C/N). The bandwidth is the ideal minimum to support the information rate. The C/N is in terms of a steady signal and Gaussian noise

---

Paper 61-1112, recommended by the AIEE Radio Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted February 16, 1961; made available for printing September 15, 1961.

J. R. Poppe and A. J. Vadasz are both with the General Electric Company, Lynchburg, Va.

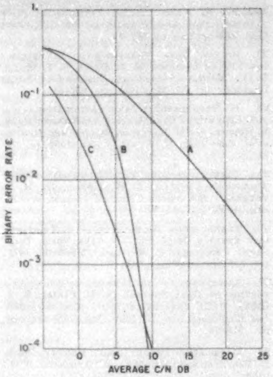


Fig. 1. Element error versus average carrier to noise ratio

A—Rayleigh fading
B—No fade
C—Rayleigh quadruple with maximal ratio diversity

over a long period of time. Curve *A* represents the performance under slow Rayleigh fading conditions with C/N in terms of the long term median. Curve *C* is the performance obtained from four uncorrelated Rayleigh-fading signals which are combined in a maximal ratio fashion.

Although theoretical, these curves which are based on the work of Montgomery[3] indicate:

1. The relative efficiency of binary detection of coded signals, as compared to analog voice
2. The undesirability of Rayleigh fading when low error rates are desired
3. The substantial benefits of diversity combining.

Theoretically a nonfading carrier should deliver a binary error rate of 0.1% with a C/N of 8 db (decibels) (Curve *B*). In terms of 60 words per minute (wpm) Teletype, this is obtained with a voice channel C/N of −14 db which is considerably below voice communications quality.

Even though these curves are idealized, important benefits are still available under practical less optimistic conditions as shown in Fig. 2. The plots are in terms of 100-wpm Teletype and relate character error rate to median C/N. They are based on practical results obtained from several installations.[3–6] The median C/N has been normalized to a 1-kc bandwidth for easy comparison. Note the substantial improvement that dual switch diversity provides over a nondiversity system. Also note the improvement that constant ratio (and other) error checking can provide.

These low error rates and the significant improvements which diversity can provide are the clues to the benefits which can be realized from digitalized tropo-scatter.

Application to Tropospheric Scatter

It should be pointed out that our work has largely been in terms of relatively low data rates and binary detection. However, this is not the only area where coding techniques have applicability to tropo-scatter. We will undoubtedly see multilevel or multistate coding coupled with bandspreading, which offer many interesting possibilities. But available technology and hardware have been favoring the early application of binary schemes and hence this discussion will be applied primarily to that area. In applications requiring relatively low data rates, the prospect of very narrow radio-frequency bandwidths, and, as stated earlier, bandwidths just adequate to support the information rate, come to one's mind.

These narrow bandwidths are only modestly beneficial since, in practical systems with post-detection filtering, the error rate is only partly dependent on predetection bandwidth.[7] This is illustrated in Fig. 3 where deviation and predetection bandwidth have been adjusted together. Performance is plotted as a function of signal-to-noise ratio (S/N) in a normalized 1-kc channel. It is important not to think of FSK as a form of frequency modulation in the voice sense. It is a different phenomenon.

The curves presented thus far are applicable to a start-stop type of Teletype channel. It should be noted, however, that the start-stop feature is not desirable for radio circuits because of the disproportionate error-producing capabilities of a mutilated start or stop pulse. Furthermore, printers have the characteristic of losing synchronism over a sequence of characters once a stop pulse is missed. The actual difference between character error rate and binary error rate ratios for start-stop and synchronous types of operation becomes a function of the decoding device and the fading characteristics of the radio transmission path. Generally speaking, the ratio between synchronous and start-stop error rates may be assumed to be between two and five to one. Thus the behavior of a four-channel synchronized time-division multipler system may be considered substantially equivalent to a single start-stop channel including adjustments for pulse speed, synchronization, and removal of start and stop pulses.

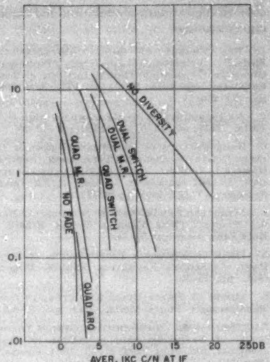


Fig. 2. 100-wpm Rayleigh fading Teletype performance

For the direct transmission of frequency shift Teletype over a UHF carrier the combination of coding, binary detection, and efficient combining permits the use of relatively low gain antennas and low power transmitter over substantial distances.

System Configurations

Major system parameters may be traded off against one another as shown in Fig. 4. A dual diversity system is assumed and a maximum character error rate of 1 in a 1,000 which is not exceeded for 99% of the annual hours in temperate

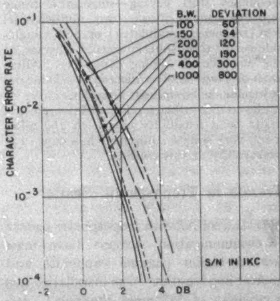


Fig. 3. Frequency shift keyed error rates for various deviations

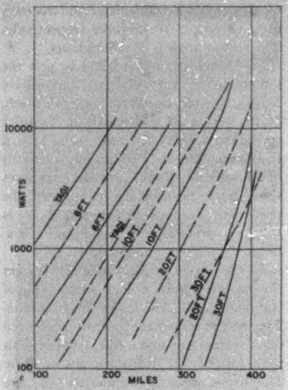


Fig. 4 (left). Required transmitter power versus miles for given antenna of 100-wpm Teletype

Solid line—900 mc
Broken line—400 mc

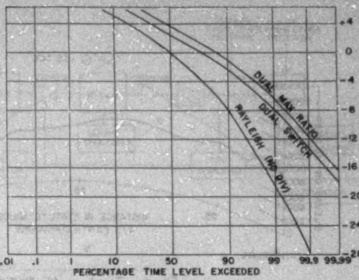


Fig. 5 (right). Dual diversity distributions

zone service is obtained. Receiver head ends are of the conventional type. A 9.5-db reserve factor for a 95% service probability is also included. Path calculations were made using the method recommended by the National Bureau of Standards.[8,9] Note that, for a given antenna diameter, a lower transmitter power level is required for 900 megacycles (mc) than for 400 mc (up to about 400 miles in hop length). A given antenna gain configuration, however, such as the 18-db Yagi array which is included in the curves, requires a lower power level at 400 mc than at 900 mc. At distances greater than 400 miles, the antenna curves cross due to the emergence of antenna-to-medium coupling losses which penalize the higher frequencies due to the narrower beamwidths realized. For nondiversity service, a transmitter power increase factor of 8 is required whereas quadruple diversity would permit a reduction of 4.5 times.

Distances are based on smooth earth profiles and do not assume geographical enhancement or penalty. This is probably a pessimistic assumption in comparison with actual tropo-scatter installations. Because of the substantial cost advantages realized by locating stations on sites featuring superior horizons, the tropo-scatter engineer has tended to put antennas on hilltops and cliffs. For the type of service described in this paper, this may not always be desirable since system margin is generally abundant and cheap.

There are a number of equipment considerations that require close scrutiny before applying these principles to hardware. Depending upon the specific modulation scheme used, design of the basic transmission circuit should be used to match the characteristics of the information channel to the properties of the propagation medium.

The successful detection of digital information streams requires the maintenance of at least a very modest C/N essentially all of the time. Maintenance of signal level substantially above this threshold is usually useless; anything below this threshold is detrimental. Thus the system designer attempts to realize this threshold point at as low a carrier level as possible. This leads to a significant departure from system designs typical of voice in which relatively high average S/N are required and occasional dips far below threshold are tolerable. (Regeneration or reshaping of pulses at repeater stations also permits satisfactory operation of digital transmission circuits through relatively high average noise levels.)

An additional tool that could substantially extend the usability of this method is time diversity. This hinges on the premise that the information capacity of the tropo-scatter signal is not effectively utilized when the system is engineered to work with the very low signal levels which prevail for only short intervals of time. In Fig. 5, it can be seen that, under Rayleigh fading conditions, a threshold improvement of 10 db can be achieved at a loss of less than 1% of transmission time by relaxing the reliability requirement from 99.9% to 99%. This sacrifice is well worth the improvement. In space or frequency diversity systems, the rewards of this manipulation are more modest, but in many cases are still remunerative. The digital tropo-scatter engineer thus has a very valuable tool in reserve to solve many system problems.

Bidirectional error-checking systems often already possess the basic elements of time diversity operation. However, a much simpler version using signal level as the transmission criterion instead of, or in addition to, constant mark/space ratio or parity bits offers many interesting possibilities. Here there is a similarity to the type of time diversity used in meteor burst systems.[10] The main difference in applicability to the two propagation methods is that in the case of meteor burst systems, time diversity is the basic condition for information transfer whereas with digital tropo-scatter, it is an optional tool.

Is this system an improvement or replacement for existing tropo-scatter designs or for high-frequency circuits? The answer is not a simple one. It is preferable to think of it not as a replacement but rather as a new tool to fulfill requirements for reliable data, Teletype, and similar types of transmissions.

Verification Program

In November, 1959, the Communication Products Department of the General Electric Company established a test circuit from Arlington, Virginia, to Lynchburg, Virginia. This 152-mile system was built to prove the feasibility of the theoretical conclusions given in the preceding paragraphs and also to gain experience in propagation, construction, and operation necessary to serve as a basis for the proposed design of narrow-band low-power tropo-scatter systems. With a transmitter power of 250 watts and dual diversity receivers with selection combining, successful results have been obtained with FSK 60-wpm start-stop Teletype. The calculated median signal loss for this path in time block II (November–April, 1 pm–6 pm) was 202 db; these measurements showed an actual loss 207 db.

In obtaining information for further equipment design through establishment of a test link, several alternatives were open. To keep decisions as simple as possible, primary emphasis was placed on frequency. The choice was made by

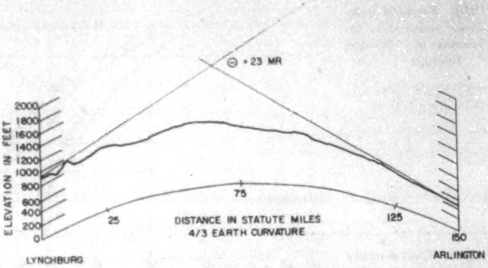


Fig. 6. Terrain profile K12XFM

Distance 152 miles

considering the aspects of hardware technology, spectrum availability, and the properties of the propagation mechanism under consideration, and their interrelationships. Low-frequency operation was selected and some of the determining factors follow:

1. For a given configuration of fixed array antennas (Yagis, dipoles, or arrays), the gain is about the same over a wide range of frequencies if actual physical size is neglected. Therefore, better system performance is realized at lower frequencies with their lower space loss.
2. Frequency stability requirements are easier at the lower frequencies.
3. On very long paths tropospheric-scatter propagation is more favorable for the lower frequencies.
4. Components are more available at lower frequencies and generally are cheaper. This is especially true of transmitters.
5. On very short paths, the possibility of using the diffraction mode is particularly interesting for cases with poor horizons such as those encountered with "hardened" sites. Lower frequencies are favored on paths such as these.
6. In mountainous areas where knife-edge enhancement is possible, the lower frequencies are preferable.

Factors favoring the choice of higher frequencies include the possible use of fixed area antennas, particularly small antennas such as horns and paraboloids. If limited to a given antenna diameter, higher frequencies are favored within certain frequency and path length limitations; see Fig. 4. On short paths where small antennas can be used and the diffraction case is not of interest, system loss may be less because of this consideration. Spectrum space, generally speaking, is more available at the higher frequencies.

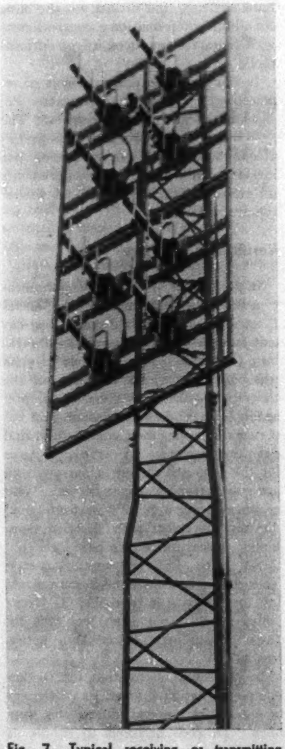


Fig. 7. Typical receiving or transmitting antennas

Experimental Path and Equipment

The profile of the 152-mile path is shown in Fig. 6. The transmitter has a clear view of the ratio horizon 11.5 miles away. The two receiving antennas are located on the roof of the General Electric plant in Lynchburg and are separated by a distance of 200 feet. The horizon angle of elevation for transmitter and receiver are −8 and +10 minutes respectively. The scatter angle is 23 milliradians.

To reduce complexity and to test the low-frequency case, relatively inexpensive Yagi arrays shown in Fig. 7 weighing only 50 lbs were used. Mounted on dual flat screens measuring 3×4 feet, they have a free space gain of 17 db and beamwidths of 25 degrees. For all practical purposes these large beamwidths eliminated the precarious antenna orientation which is so often necessary in tropospheric systems.

In the design of any diversity system the engineer must decide what type of combining will yield the best results. The test system has approximately 225 db of system gain and a calculated median path loss of 202 db, a margin of some 23 db. Careful examination of the distribution curves for a dual diversity system shows that differences between combining schemes is quite small. For example, the 99.99% exceeded level of the selection-diversity curve is almost 20 db above the Rayleigh curve, while the maximal-ratio curve is only 1.4 db better. For a dual diversity system a choice would be made between the two only if the last decibel were crucial and even then it is doubtful that this improvement could be realized in an operating system. On this basis, selection combining was chosen. For a system operating with higher orders

Fig. 8. Thin route tropospheric transmitter

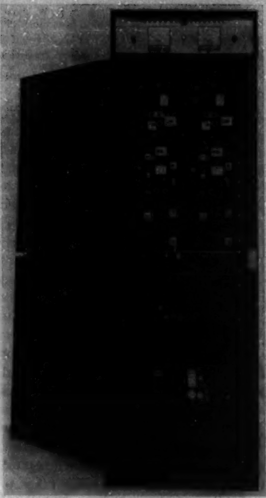


Fig. 9. Thin route tropospheric dual diversity receiver

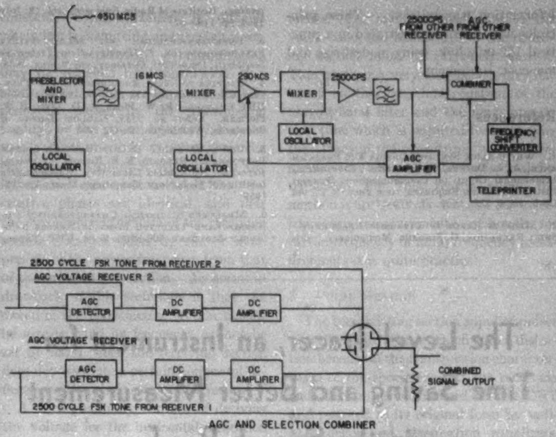

450 MCS
PRESELECTOR AND MIXER
LOCAL OSCILLATOR
16 MCS
MIXER
LOCAL OSCILLATOR
290 KCS
MIXER
LOCAL OSCILLATOR
2500 CPS
2500 CPS FROM OTHER RECEIVER
AGC FROM OTHER RECEIVER
COMBINER
AGC AMPLIFIER
FREQUENCY SHIFT CONVERTER
TELEPRINTER
2500 CYCLE FSK TONE FROM RECEIVER 2
AGC VOLTAGE RECEIVER 2
AGC DETECTOR
DC AMPLIFIER
DC AMPLIFIER
AGC VOLTAGE RECEIVER 1
AGC DETECTOR
DC AMPLIFIER
DC AMPLIFIER
2500 CYCLE FSK TONE FROM RECEIVER 1
COMBINED SIGNAL OUTPUT
AGC AND SELECTION COMBINER

Fig. 10. Block diagram of single receiver and combiner circuitry

of diversity, a different choice may be advisable.

In establishing the test link, the primary goal was to obtain experience as quickly as possible. Therefore, the equipment used to make these tests was accumulated from various other projects within the plant.

The exciter and power amplifier in Fig. 8 are a standard General Electric station made for the mobile services, using a type 4CX300 output tube. The frequency source is a high stability oscillator of the type currently used on the AN/FRC-47 equipment. It has a short term stability at 12.5 mc of 1 part in 100 million per day. This high degree of stability and day-to-day maintenance of the equipment eliminated the need for any afc (automatic frequency control) at the receiver. (Future equipment designs will be furnished with a long term afc to eliminate this day-to-day maintenance and assure unattended operation.) The receiver (Fig. 9) input is connected to a coaxial preselector-mixer cavity that provides approximately 20-db image rejection and an over-all noise figure [mixer plus if (intermediate frequency)] of 6 db. The first if is 16 mc, the first local oscillator frequency multiplying factor is 36 and the basic oscillator frequency 12.055 mc. There are three frequency conversions from the signal frequency to 2,500 cycles, the last if; see Fig. 10. These are 16 mc, 290 kc, and finally 2,500 cycles per second. All of the frequency conversion oscillators are identical to the high stability oscillator in the transmitter. The 2,500-cycle translation oscillator operates without temperature stabilization. The receiver 3-db bandwidth is 500-cycles determined by an inductance-capacitance filter with center frequency at 2,500 cycles. All combining and agc (automatic gain control) detection is done after the 2,500 cycles filter. The combiner tube is a dual triode with a common cathode. The agc voltage of each receiver is passed through a differential amplifier and applied to the grid of each half of the tube. The signal from each receiver is also tied to the appropriate grid of the combiner. When connected in this way, that side of the tube having the lower agc voltage is cut off and the signal from the corresponding receiver does not contribute to the output. The switching circuitry has a switching time of 0.1 millisecond and switches when a differential of 0.1 db exists between the two agc voltages. For the bit rates used, this switching time has been relatively short and no errors due to switching have resulted even at low signal-noise ratios.

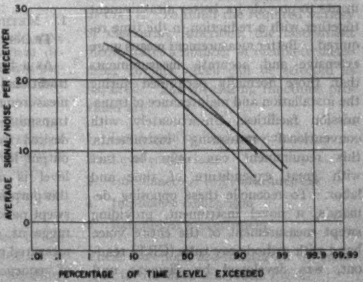


Fig. 11. Distribution received signal levels from November to February

Typical signal level distributions for the winter months are shown in Fig. 11. Propagation during the summer months has been even better. Error rates as low as 0.001% are achieved practically every day.

Conclusions

Direct application of digital modulation techniques to a radio frequency carrier using the tropospheric scatter mode of propagation (FPTS) portends many benefits, especially in range, savings in plant investment, and reliability. When coupled with coding techniques, it becomes a reliable and efficient means information transmission. These principles have been demonstrated in a practical 152-mile link, using modest-size and low-cost equipment.

References

1. White Alice System—Design and Performance, **A. L. Durkee, D. Metcalfe.** *Transactions*, Professional Group on Communications Systems, Institute of Radio Engineers, New York, N.Y., vol. CS-7, no. 4, Dec. 1959, pp. 272–77.

2. Message Error in Diversity Frequency—Shift Reception, **G. Franklin Montgomery.** *Proceedings*, Institute of Radio Engineers, vol. 42, July 1954, pp. 1184–87.

3. Funkfernschreiben mit selbsttaetiger Fehlerkorrektur, **F. Hennig.** *Neue Technische Zeitschrift*, Düsseldorf, Germany, vol. 9, Aug. 1958.

4. Carrier-to-Noise Requirements for Teletype Communication via Non-fading or Fading UHF Carriers, **R. W. Plush, A. D. Watt, E. F. Florman.** *Report no. 5585*, National Bureau of Standards, Washington, D. C., 1960.

5. The AN/FRC-47 (XD-1) Mock-up System Performance Results, **B. E. Nichols.** *Technical Report no. 203*, Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., Apr. 23, 1959.

6. Measured Statistical Characteristics and Narrowband Teletype Message Errors on a Single Sideband 600-Mile-Long UHF Tropospheric Radio Link, **R. F. Florman, R. W. Plush.** *Report no. 6019*, National Bureau of Standards.

7. Fehlerhaeufigkeit von A- und F-Telegraphie Uebertragungs—Systemen insbesondeu bei Weissem Ranchen, **Horst Beger.** *Telefunken Zeitung*, Berlin, Germany, vol. 29, Dec. 1956, pp. 245–55.

8. Prediction of the Cumulative Distribution with Time of Ground Wave and Trospospheric Wave Transmission Loss, **P. L. Rice, A. G. Longley, K. A. Norton.** *Report no. 5582*, National Bureau of Standards, 1958.

9. Predicting the Performance of Long Distance Tropospheric Communications Circuits, **A. P. Barsis, K. A. Norton, P. C. Rice.** *Report no. 6032*, National Bureau of Standards.

10. Meteor Burst Systems. *Proceedings*, Institute of Radio Engineers, vol. 45, Dec. 1957.

The Level Tracer, an Instrument for Time Saving and Better Measurement in the Speech Band

MARTIN NIEDEREDER
NONMEMBER AIEE

BENJAMIN T. SANDERS, JR.
ASSOCIATE MEMBER AIEE

Summary. The development and construction of a visual voice-frequency transmission-measuring set is described, and a short summary of its applications is included. The instrument provides for the display of measurements of level, impedance, return loss, impedance comparison and level comparison on the face of a CRT (cathode-ray tube) as a function of frequency. The values in db (decibels) or ohms are read directly from a calibrated graticle. The instrument described is the latest of a series whose initial development was begun in 1941.

TELEPHONE connections in growing number, the increased demands on transmission performance, and the increasingly serious shortage of skilled personnel have strongly emphasized the need for communications-measuring instruments to provide for better measurement together with a reduction in the time required. Better measurement means more extensive and accurate measurements than those normally performed during the installation and maintenance of transmission facilities. Unfortunately with conventional measuring instruments, this requirement can only be met with great expenditure of time and labor. To reconcile these opposing demands, a novel instrument, providing swept measurement of the entire voice band with cathode-ray tube (CRT) readout, was developed by Siemens und Halske AG and marketed under the name Level Tracer. The development of the technique of measurement exemplified by this instrument dates from 1941 and an earlier version is described in a previously published paper.[1]

The basic development problems and their solution as well as the extensive measuring capabilities and some practical applications of the latest version of this instrument are surveyed herein.

Background and Development Problems

A number of possibilities exist for the technical layout of a visual transmission measuring set, whose advantages and disadvantages will be briefly considered below.

1. Method with Periodically Varied Transmitting Frequency

As a first solution, automation of the traditional technique of transmission measurement suggests itself, where the transmitted frequency is varied over the desired frequency range with constant output level, and the resultant received level is measured and recorded. For this purpose the transmitting frequency is swept periodically through the frequency range at constant amplitude, and the measured result is displayed as a function of frequency on the face of a CRT.

(a) *Transmitting section*

It is well known that a frequency band of more than four octaves can be covered without band switching only by means of the beat frequency generator principle, where the required frequency band is derived as the difference of two considerably higher oscillator frequencies, of which one is fixed while the other is varied through the range Δf. This Δf is equal to the desired frequency band, and referred to the higher oscillator frequencies, it is relatively narrow. The choice of oscillator frequencies depends upon the ease with which the desired difference frequency band may be achieved, but it is limited on the other hand by the point at which the frequency stability of the two oscillators becomes difficult to maintain. A third consideration in the selection of the oscillator frequencies is the means by which the detuning is to be realized. Swept-band operation is possible electronically with reactance tubes, magnetically-biased coils, voltage-dependent capacitances, etc., or mechanically with a rotary variable capacitor.

A rotary variable capacitor establishes a definite initial capacitance, and permits a wide and highly stable capacitance variation to be readily attained. Relatively low oscillator frequencies can thus be chosen, and after simple standardization of the calibration, a well-defined difference-frequency band is obtained. The low value of the oscillator frequencies permits excellent generator stability with relatively moderate expense. The sweep generator may also be used as

Paper 61-1072, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted June 16, 1961; made available for printing September 19, 1961.

Martin Niedereder is with Siemens und Halske AG, Munich, Germany, and Benjamin T. Sanders, Jr., is with Sanders and Associates, Inc., Sante Fe, N. Mex.

an ordinary variable frequency oscillator, since the transmitted frequency can be selected manually from the calibrated scale of the variable capacitor.

Electronic sweep methods require more complicated circuitry for comparable stability, which results in greater expense and correspondingly increased maintenance. The only significant advantage of an electronic sweep is its higher possible sweep rate. Inasmuch as the upper limit for the sweep rate is imposed by the electrical properties of the specimen and the detector, no significant advantage may be derived from this characteristic within the voice band. A principal disadvantage of conventional electronic sweep methods is that manual selection of the transmitted frequency is not possible. It may be readily determined from these facts that the mechanical sweep technique provides a number of advantages not afforded by the electronic method. The problem of limiting the sweep rate is discussed below.

(b) *Receiving section*

For indication of the measured results, the amplitude and frequency values can be correlated in various ways for display on the CRT. Modern CRT's possess sufficient linearity and freedom from distortion that direct calibration of the screen graticle is possible if the supply voltages are regulated. The measured results can be read in the same manner as from a pointer instrument, where the vertical axis of the graticle is calibrated in dbs and ohms. With a flat-face CRT, the parallax error can readily be held below 1%. Unlike a pointer instrument with a single dimension, the CRT permits two-dimensional display. The measured results can, therefore, be displayed as a function of frequency. This requires a frequency dependent horizontal deflection of the electron beam, which can be accomplished in a number of ways.

The simplest method is that of controlling the horizontal deflection directly by the same mechanical or electrical process that controls the frequency sweep of the transmitting section. With this method, however, the transmitter and receiver cannot be separated unless a means exists for transmitting the horizontal signal over a separate path.

A second well-known method derives the horizontal deflection of the CRT from a sawtooth generator. This generator must be synchronized with the transmitting section, and the time functions of the frequency sweep of the transmitter and the horizontal deflection of the receiver must be exactly co-ordinated and constant. Synchronization is difficult where the transmitting and receiving sections are separated. Synchronization may be achieved by transmission of a start signal using a frequency that is positively passed by the transmission facility, or the power frequency may be used, where there is coupling between the power distribution systems. In the latter case, care must be taken that the relative phases are identical, and that the synchronizing system of the receiving section does not differ by one or more periods of the power frequency from that of the transmitting section. An essential drawback of this technique is that the sweep of the transmitting section cannot be arrested and its frequency cannot be set by hand, since the receiving section does not associate a certain horizontal deflection with a given received frequency.

A third possibility consists of deriving the voltage for the horizontal deflection of the electron beam from a frequency meter incorporated in the receiver. The received signal is thus associated with a fixed horizontal deflection that corresponds to its frequency. Independence of the transmitting section is thus obtained and the transmitted frequency may vary with time in any desired manner. No synchronization between transmitting and receiving sections is required, and the horizontal axis of the graticle may be calibrated directly in frequency.

2. METHOD WITH CONTINUOUS TRANSMITTED SPECTRUM

With the method discussed above, the transmitting frequency is varied periodically; and the received signal is detected on a wideband basis. A second method reverses this process. The transmitting section supplies a wideband spectrum, i.e., a random noise source with flat amplitude spectrum which is applied to the transmission facility, while at the receiving end a selective detector periodically scans the incoming frequency spectrum. This method offers three advantages:

(a) The specimen is always in a steady-state condition as opposed to the method described previously, where the transient performance of the specimen must be taken into account.

(b) No synchronization is required between transmitting and receiving sections since the frequency dependent element is contained in the receiving section only.

(c) Manual measurements at selected frequencies are possible.

These advantages are opposed by an essential drawback to the application of this method to voice-frequency measurements. For sufficient resolution in the voice band, the bandwidth of the receiving section must be extremely narrow. This involves a very long measuring time based on the long-time constant of the narrow-band filter and the long integration time which is required for averaging the transmitted spectrum whose frequencies are assumed to have a normal distribution. For this reason alone the method is not generally suitable for voice-frequency measurements, although for carrier frequency or radio-frequency measurements it is quite practical.

3. PULSE METHOD

The transmitting section supplies pulses of well-defined shape which suffer distortion because of the transmission characteristics of the specimen. At the receiving end, this pulse is observed on the CRT and restored to its original form by variable delay and attenuation equalizers. This method has proved extremely satisfactory for conditions where accurate determination of both attenuation and delay distortion are of primary importance. Because of the complexity of the apparatus required, the application of this technique is not advisable for general transmission measurements in the field by unskilled personnel, nor is the expense justified for measurements of attenuation alone.

Solutions and Further Considerations

Following comparison of the methods outlined and consideration of the possible applications of the instrument, it was decided that the following arrangement was most desirable, and subsequent developments and improvements have continued along these lines.

The level tracer is a complete measuring setup consisting of transmitting and receiving sections mounted in a portable case. It traces the measured quantity as a curve on the screen of the CRT. A panel view of the instrument is shown in Fig. 1. To fulfill the required variety of measuring functions, the instrument incorporates a number of complete measuring circuits which can be selected as required. The measuring functions provided are:

1. Level (gain or loss) measurement
2. Level comparison
3. Impedance ($R+jX$) measurement
4. Impedance comparison
5. Return-loss measurement

These functions are illustrated by reference to the block schematic Fig. 2.

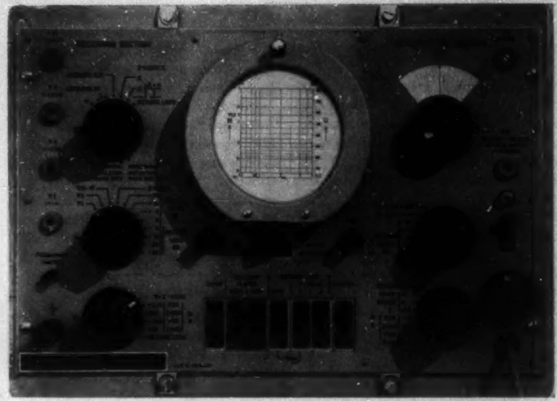


Fig. 1. Front panel view

TRANSMITTING SECTION

The frequency is periodically varied from 200 cps (cycles per second) to 4 kc (kilocycles) by a revolving variable capacitor driven by a synchronous motor. The motor can be switched off to permit the frequency to be set manually. In addition to providing swept-band operation, this method permits accurate investigation of specific points of interest within the transmission band. The variable capacitor runs on ball bearings and in conjunction with the synchronous motor provides reliable operation without noticeable wear. The first motor-driven capacitor assemblies of this kind, which were incorporated in earlier models of the level tracer, have given many years of completely trouble-free operation.

RECEIVING SECTION

The amplitude and frequency of the incoming signal are measured separately by a level meter and a frequency meter which provide the vertical and horizontal deflection voltages for the electron beam.

With a periodically varied transmitting frequency, the measured result y is traced on the screen as a function of frequency x. This method requires no synchronization between the transmitting and receiving sections, and measurements are possible over facilities of any length. The transmitting and receiving sections are calibrated by means of internal standards, and the measured results may be read directly from the attenuators and the graticule of the CRT. The y axis of the graticule is calibrated in db on the left side and ohms on the right side. The x axis of the graticule is calibrated directly in frequency.

SWEEP FREQUENCY

Before proceeding further with the description of the level tracer, some observations concerning the advisable time function of the frequency sweep and the maximum possible sweep rate are in order. These factors are limited by the dynamic behavior of the specimen and the receiving section of the measuring instrument. A primary task of the development engineer was to design the instrument so that the sweep rate is limited by the dynamic behavior of the specimen alone. It is essential that swept-band measurements provide results which are identical to the values obtained with point-by-point measurement. With point-by-point measurements the specimen is in a steady state, and the results are governed by the static performance of the specimen. A swept-band measuring method is by its very nature a dynamic method, and as a result the dynamic transfer constant of the specimen must be considered. If the attenuation of a specimen is measured with a periodically varied measuring frequency, a result is obtained, with minor deviations from the steady state condition, which is stated in the literature by the following approximation.[2]

$$a_d = a_0 + \frac{1}{2}\,\frac{d\omega}{dt}\left(\frac{d\tau}{d\omega} - 2\tau\,\frac{da}{d\omega}\right) \qquad (1)$$

where

a_d is the swept band attenuation constant
a_0 is the static attenuation constant
ω is the instantaneous angular frequency
$\tau = \frac{db}{d\omega}$ is the envelope delay (variation of the phase constant with frequency)

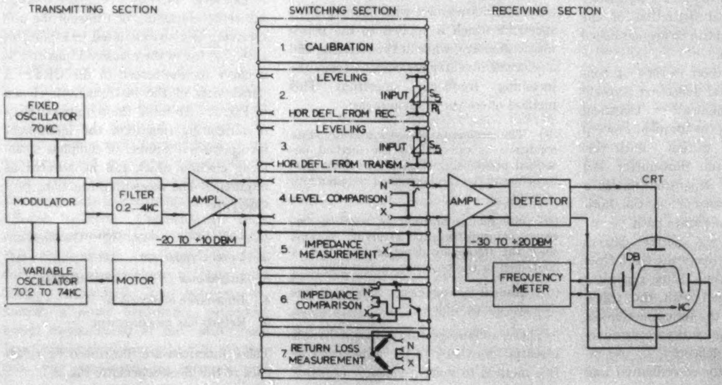

TRANSMITTING SECTION
SWITCHING SECTION
RECEIVING SECTION
FIXED OSCILLATOR 70 KC
MODULATOR
FILTER 0.2–4KC
AMPL.
VARIABLE OSCILLATOR 70.2 TO 74KC
MOTOR
−20 TO +10 DBM
1. CALIBRATION
LEVELING
2.
INPUT
S_5
R_1
HOR. DEFL. FROM REC.
LEVELING
3.
INPUT
S_5
HOR. DEFL. FROM TRANSM.
4. LEVEL COMPARISON
N
X
5. IMPEDANCE MEASUREMENT
X
6. IMPEDANCE COMPARISON
N
X
7. RETURN LOSS MEASUREMENT
N
X
AMPL.
DETECTOR
−30 TO +20 DBM
FREQUENCY METER
CRT
DB
KC

Fig. 2. Block schematic

Equation 1 shows that deviations from the steady-state condition depend upon the frequency response $da/d\omega$ of the attenuation and on that of the delay $d\tau/d\omega$. It further shows that the deviations are directly proportional to the sweep rate $d\omega/dt$. Since the dynamic characteristics of the specimen denoted by the expression in parenthesis are fixed quantity, the deviation may be reduced only by reduction of the sweep rate $d\omega/dt$. It is obvious that the direction of the deviation from the steady-state condition depends upon the algebraic sign of $d\omega/dt$, that is upon the direction of the frequency sweep. This leads to two conclusions:

1. The sweep generator should be so designed that the direction of the frequency sweep alternates, that is, the frequency band is traversed in succession from low to high and at the same rate from high to low frequencies. Any deviation from the quasi-steady-state condition makes itself known at once by the lack of coincidence between the forward and return traces on the screen.
2. With small deviations and uniform sweep rate, the average between the forward and return traces represents the correct measuring result. Since as stated by equation 1, the direction of the deviation in attenuation changes with the algebraic sign of $d\omega/dt$, the difference between the two ordinates equals the double deviation of the forward and return trace.

The foregoing establishes the necessity of providing a go and return sweep to verify that measurements are being carried out under steady-state conditions, and that the go and return must be effected at the same rate.

The time function of the frequency sweep is a matter of great importance. The measurement of wideband specimens presents no particular problem in this respect. As the bandwidth of the specimen decreases, the choice becomes more critical, as the oscillation build-up time is inversely proportional to the bandwidth of the specimen, and it is necessary for the dwell time within the transmitted band B to be at least equal to the oscillation build-up time $1/B$. The expression dwell time designates the time interval which the transmitting section requires for traversing the transmitted band B. Relatively equal specimens, in the simplest case tuned circuits of the same Q-factor but different resonant frequencies, would all be measured under the same dynamic conditions and present uniformly high or low distortion, if the time function of the sweep were chosen to vary with the square of the frequency. On the other hand, with specimens of different center frequencies but the same absolute bandwidth, a uniform sweep rate for all frequencies would be necessary. A middle path between these extremes was selected, and the sweep rate versus frequency function was chosen to be approximately proportional to frequency, corresponding to the well-known logarithmic frequency scale.

The logarithmic frequency sweep has proved most satisfactory, and it has been determined in practice that all specimens for which the level tracer is intended can be positively and accurately measured with a sweep period of approximately 1–2 seconds. With such sweep periods, specimen buildup times averaging 5 milliseconds can be neglected in practice.

The build-up time of the receiving section of the level tracer is defined by the detector time constant required for demodulation of the lowest measuring frequency (200 cps) and the time constant of the succeeding low-pass filter for the harmonics produced in the demodulation process. By suitable circuitry this build-up time has been reduced to an appropriately small value.

In earlier versions of the level tracer where little experience had been gathered concerning the sweep time required for various types of specimens, the received frequencies were translated to a higher frequency position, which reduced the build-up time of the receiving section. This permitted the instrument to be operated with sweep rates as high as 5 cps. It was determined in practice that so few specimens could be measured at the faster rates that the increased complexity and expense required was not justified.

Functions and Operation of the Instrument

The present level tracer *Rel 3 K 211 g* is the product of 20 years of research, development, and application of the swept-band measuring technique. With the exception of modifications incorporated for the specific requirements of the U.S. market, the instrument has been in use throughout the world in its present form for over 8 years and is capable of solving practically all problems encountered in voice-frequency and message channel measurement.

Design emphasis has been placed on utmost simplicity of calibration and operation of the instrument so that craft personnel can readily perform reliable measurements without extensive training. Since the full voice band is traced without a gap, irregularities of the communications facility under test are readily detected by inspection of the CRT display.

The instrument is constructed of precision components throughout whose useful life and stability are far superior to commercial grade components. Only one type of amplifier tube is used, the type *C3M*, which is a long life type capable of at least 10,000 hours of service. When replacing these tubes or the CRT, no additional action whatsoever is required inside the instrument, such as trimming, correcting, etc. Tubes can, therefore, be replaced by unskilled personnel in the field or elsewhere, should the need arise.

CALIBRATION

Calibration of the instrument is carried out in six simple steps as follows:

Calibration 1. The measuring trace is adjusted to coincide with the infinite loss line of the CRT graticle.

Calibration 2. The measuring trace is adjusted to coincide with the 0-loss line of the graticle of the CRT. For this purpose the regulated calibration voltage is applied to the input of the receiving section and its sensitivity is adjusted against this standard.

Calibration 3. The output of the transmitting section is adjusted to coincide with the 0-loss line of the CRT graticle.

Calibration 4. The frequency sweep is arrested, and the oscillator is adjusted manually to 800 cps. The beat-frequency generator is adjusted for zero beat against an internal 800-cps resonant-trap circuit by depressing and rotating the trimmer designated calibration 4.

Calibration 5. The oscillator is set to 4 kc and the measuring spot is adjusted to coincide with the 4-kc graduation of the CRT graticle.

Calibration 6. The oscillator is set to 200 cps and the measuring spot is adjusted to coincide with the 200-cps graduation of the CRT graticle.

The level measuring accuracy of the receiving section is equivalent to that of a corresponding pointer type instrument since the sensitivity of the CRT is sufficiently stabilized by regulation of the supply voltages. For calibration of the level tracer a regulated voltage supply is incorporated. Experience has shown that the calibrating source maintains the guaranteed measuring accuracy of the unit over a period of 4 years. Thereafter, it should be compared against an external standard to correct for errors of the calibrating voltage of about 1 to 2% which occur as a result of component aging.

LEVEL MEASUREMENT

The level measuring function is used most frequently for determining the overall equivalent of a transmission facility. For this purpose, the output impedance of the transmitting section and the input impedance of the receiving section must be matched to the nominal impedance of the transmission path. This is effected with a switch which selects transmitter source impedances of ≈0, 600, or 900 ohms. The receiver impedance may be selected in the same manner as 600 ohms, 900 ohms, or High. Inasmuch as the instrument is calibrated in power level dbm (decibels referred to 1 milliwatt), switching of the impedance values is accompanied by switching of the emf (electromotive force) of the transmitting section and the sensitivity of the receiving section.

The transmitting level may be controlled by means of an attenuator from −20 to +10 dbm in 5- and 10-db steps as follows: +10, +5, 0, −5, −10, −20. The full scale sensitivity of the receiving section may be adjusted over the range from −30 to +20 dbm in the following steps: +20, +10, +5, 0, −5, −10, −20, −30. Within the sweep range of 200 cps to 4 kc, the emf of the transmitting section and the sensitivity of the receiving section remain constant within ±0.2 db. The absolute received level can be read directly as a function of frequency on the calibrated graticle of the CRT, and by reference to the transmitted level, the over-all equivalent may be determined.

The ordinate of the CRT graticle is graduated logarithmically over the range from −20 to +3 db. The range of −5 to +3 db is graduated in 1-db steps and that of −5 to −20 db in steps of 5 db. In the most sensitive position of the receiving level switch, an input level of −50 dbm is easily readable. The abscissa of the graticle is approximately linear as a function of frequency, and is graduated in 100-cps divisions from 200 to 1,000 cps and in 500-cps divisions from 1,000 to 4,000 cps. With use of the maximum transmitting level of +10 dbm, the attenuation measuring range is 60 db. Conversely, gains up to 43 db may be measured without the use of an external attenuator.

With a power frequency of 60 cps, the sweep time of the transmitting section from 200 cps to 4 kc is 1.66 seconds, providing a go and return time of 3.32 seconds. This sweep rate permits measurements under quasi-steady-state conditions, while at the same time the long persistence screen of the CRT permits the observation of a continuous trace. As stated above, the go and return frequency sweeps are effected at the same rate. Therefore, lack of coincidence of the traces reveals the absence of a steady-state condition.

Where the build-up time of the specimen is so great that the steady state is not approximated, the sweep of the transmitting section should be arrested manually or by disconnection of the drive motor, and the tuning capacitor rotated slowly over the frequency band. The correct response curve of the specimen will be traced on the face of the CRT. When the sweep of the transmitting section is arrested, the indication remains on the CRT at the receiving end. The luminous spot, which in swept operation traces out the whole curve, comes to a stop at the frequency on the curve where the sweep has been stopped. The transmitted frequency may then be set by hand as with an ordinary variable-frequency oscillator, and the luminous spot will always display the received level as a function of the frequency received.

The use of a frequency meter to determine the horizonal deflection of the CRT permits the level-tracing receiver to interwork with any other signal generator within its bandwidth, and to indicate the received level as a function of frequency. This other generator may be a signal generator of fixed frequency, e.g., a 1-kc milliwatt supply. The level-tracing receiver will then display a spot on the screen which indicates the received level at the 1,000-cps mark on the frequency scale. The other generator may also be one that varies the frequency automatically in steps, e.g. 300 cps, 400 cps, etc., to 4 kc at various intervals. The level-tracing receiver will always display a spot corresponding to the frequency and level received from the generator. Finally, the other generator may also be of the swept type. The frequency band can be swept at any rate up to 1 cps, and the level-tracing receiver will plot the correct curve.

Since no synchronization between transmitter and receiver is required, the transmitting and receiving sections may be powered from different mains; the transmitting section may be operated on a 60-cps supply and the receiving section on 50-cps supply, and vice versa.

It is also possible to measure from a transmitting section in a different frequency position of a multichannel carrier system. If, for instance, a transmitting signal is fed into a carrier system in the line frequency position, measurement of the voice frequency channel response is possible with the level-tracing receiver connected to the channel demodulator.

With measurements of over-all equivalents, it has become customary to measure complete transmission paths including their dialing facilities. In this case the test connection is set up for transmission-loss measurements through the appropriate test boards in the same manner as with single-frequency measurements. If these dialing facilities utilize in-band signaling, the swept-band signal may disturb the measurement or the completed condition of the over-all connection when it sweeps through the signaling frequency. In swept operation, this annoying fact may differ depending upon the type of signaling employed. Either the signaling facility will split the circuit for the period the transmitting section traverses the critical frequency range, in which case a gap will appear in the trace; or if this period is long enough or the signaling system is not properly aligned, an on-hook condition will result and the test circuit will be released. In this case it will be necessary to dial time and again since the circuit will be released on each sweep.

To prevent this a band-stop filter which suppresses the critical frequency band can be activated in the transmitting section. This is initiated by a cam on the axle of the variable capacitor which actuates a contact to remove the disturbing frequency range from the swept band. The position of the cam on the axle, and accordingly the location of the suppressed band, can be set to match to the signaling frequency of the system. The bandwidth suppressed is approx. ± 150 cps. The suppression is effected by shifting the frequency without disconnecting the signal level which prevents build-up transients after reactivation. As presently furnished the level tracer is adjusted to blank 2,600 cps ± 150 cps.

The band shift is easily realized because of the means of horizontal deflection employed in the level tracer. The electron beam traces out the display up to the critical frequency, jumps back about 300 cps, corresponding to the shift of the transmitting frequency, and proceeds once more to the critical frequency. Thereafter, the beam jumps to the upper limit of the suppressed frequency band and completes the remainder of the trace. On the return sweep the process is reversed. This facility permits level and return-loss measurements over telephone circuits with any type of in-band signaling system, and it may also be adjusted for use with out-of-band signaling systems if required.

Apart from measurements of over-all equivalents using the switched internal terminations, it is also possible to measure the levels of terminated systems In the setting for high impedance input, measurements on a bridging basis

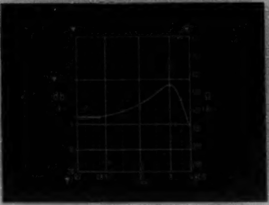


Fig. 3. Display with frequency marker at 3 kc

are possible with the level-tracing receiver at different points along a transmission path. The input impedance of the receiving section is sufficiently high that the bridging error is less than 1% for connection to a system whose impedance equals 600 or 900 ohms. For indication of the power level, the voltage sensitivity of the receiving section is switched depending upon the impedance of the system being bridged. It is thus possible to measure the power level across system terminations of 600 ohms or 900 ohms on a high-impedance bridging basis without introducing an additional error of more than 1%.

The output of the transmitting section and the input of the receiving section are floating and balanced to prevent measuring errors due to longitudinal unbalance of the facility under test. The balance to unbalance ratio of both output and input exceeds 40 db. This insures that longitudinal voltages are attenuated by a factor of 100 and can disturb the measurement by only 1% of their magnitude. This value has always proved sufficient for this class of measurements.

Where the input and the output of the specimen under test are available at the same location, the transmitting section may be switched to manual operation for closer inspection of points of particular interest on the measured curve. If the transmitting and receiving sections are at different locations and use is made of the transmitting section of one instrument and of the receiving section of another, communication by order wire would be necessary if the same method were used. For this reason, a frequency marker has been added to the display. For the generation of this frequency marker, the unused transmitting section of the receiving level tracer is used. By means of a push-button switch, the frequency meter circuit is transferred from the receiving to the transmitting section, so that a vertical luminous line is traced out at the point on the abscissa corresponding to the transmitted frequency. For this purpose, the transmitting section is set to the frequency of interest, the push button is depressed, and the value of the ordinate is read at the intersection of the trace and the frequency marker. Alternatively, the frequency marker may be moved along the trace so that frequencies of interest can be read accurately from the frequency scale of the transmitting section. This is illustrated in Fig. 3.

The instrument is also sensitive to interfering voltages which may appear as a result of power induction, crosstalk noise, etc. The magnitude of the disturbing voltage may be estimated from the deviation of the trace, and under certain conditions the frequency of the interfering voltage may be readily determined. Fig. 4 illustrates the effect of 50-cps power induction. The presence of disturbing voltages of these types is much more readily apparent when displayed on a CRT than when indicated on a pointer instrument. The integrating characteristics of the pointer instrument will frequently fail to reveal the presence of such disturbing voltages, whereas the CRT, by virtue of its rapid writing rate, will readily yield the information mentioned.

Impedance Measurement

As mentioned in the introduction, the level tracer is designed for measurements of impedance and return loss as well as for level measurements. With the exception of balancing networks for return loss measurements, no external accessories are required.

The magnitude of the impedance Z is read directly on a scale calibrated in ohms at the right side of the CRT graticle. It is graduated from 100 to 800 ohms, and the measuring range can be expanded downward and upward by the ratios 0.5 and 2 with a multiplier switch. No additional calibration is required for the measurement of impedance with the level tracer. The control knobs are set to the positions marked Z Measurement and the specimen is connected to the jack designated Z Meas. X. The impedance of the specimen as a function of frequency is then displayed directly on the face of the CRT. For this measurement the generator is connected to present a high-source impedance, i.e., constant-current generator. A current is passed through the specimen whose magnitude is effectively independent of the impedance under test. The voltage drop across the specimen is then proportional to the magnitude of the impedance $(R+jX)$. In the normal connection the voltage across the specimen equals that voltage level which is obtained by referring the impedance curve to the level scale of the graticle. At 600 ohms, for example, the level is 0 db. The level is independent of the setting of the impedance multiplier, and the level across the specimen can always be read directly from the level scale. A level of about 0 db will ordinarily be permissible. In those cases where a lower level is required, the level across the specimen can be reduced in steps by 5, 10, 20, or 30 db without changing the impedance measuring range. To accomplish this the attenuators of the transmitting and receiving sections are moved in steps in the same sense. For example, the transmitting level is reduced by 10 db while the sensitivity of the receiving section is increased by 10 db. In addition the impedance-measuring range may be extended with some sacrifice in accuracy by reducing the transmitted level in 5-db steps while maintaining a constant received level. The same result may be achieved by maintaining a constant transmitting level and changing the sensitivity of the receiving section in 5-db steps. The impedance measuring range is increased by an additional multiplier of 1.78 for each 5-db step. This feature is of great value when facilities containing severe impedance irregularities must be examined and when it is desired to bring the entire display on scale.

The impedance-measuring feature is particularly well adapted to the measurement of impedance characteristics of exchange and toll cables and to the location of impedance irregularities. Impedance unbalances may be detected, using the level measuring function of the instrument, in conjunction with a suitable external network. The location of loading faults and line faults by the analysis of impedance and impedance-unbalance curves has long been recognized as one of the most desirable techniques of fault location, but the time-consuming nature of accomplishing these measurements by conventional methods has prevented their application on other than a sampling basis, or where definite faults must be located regardless of cost. The impedance-meas-

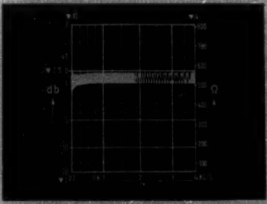


Fig. 4. 50-cps disturbing voltage

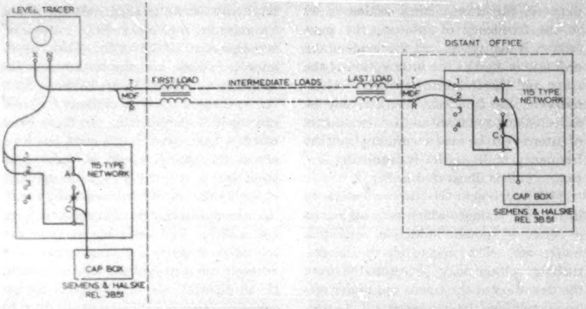


Fig. 5. Circuit for cable acceptance tests, impedance measurement, impedance comparison, and return loss

uring circuit permits the inspection of impedance characteristics on a serial basis with immediate indication of whether the facility is correctly loaded and terminated or not. Improper end section build-out of toll cables is readily discernible at distances of 30–40 miles. Fault location is easily accomplished by means of Kuepfmueller's method, which requires determination of the frequencies at which the impedance peaks and valleys occur.[3] For this purpose the sweep drive is shut off, and the transmitting section is tuned manually to the individual impedance peaks and valleys whose frequencies are read directly from the scale of the oscillator.

The technique which is referred to as "the approximate method of locating circuit faults" is immediately applicable to all impedance measurements of transmission facilities with this instrument. The distance to the fault is determined on the basis of the equation

$$d = V/2\Delta f \qquad (2)$$

where

d = distance to the fault
V = velocity of propagation
Δf = frequency difference between adjacent impedance peaks

If the exact makeup of the facility is known, the technique which is referred to as "the refined method of locating circuit faults" may be applied with equal ease and the fault isolated within 1/4 of a loading section in most cases.

Return Loss Measurement

A third measuring circuit incorporated in the instrument permits direct measurements of the reflection coefficient. The logarithm of the reflection coefficient, designated Return Loss, is indicated directly in decibels. The measuring range extends from 0 to 40 db corresponding to reflection coefficients of 100% to 1%. A 50-db return loss can be readily determined with reduced accuracy. The normal level scale of 0 to −40 db, the sum of the attenuator and graticle readings, is used for this measurement, but the result is expressed without regard to sign. The measuring circuit is a bridge incorporating a hybrid coil which is mounted within the instrument. The trans-hybrid loss of the coil is automatically compensated for when the instrument switches are operated to the Return Loss measuring positions. In a balanced condition of the bridge, that is with low-reflection coefficients, the level across the specimen is 0 db. With high-reflection coefficients the level across the specimen may increase by as much as 6 db. If the need arises, it is possible to decrease the level across the specimens in steps of 5 db.

The incorporation of an internal hybrid coil eliminates the use of external test hybrids under most conditions and considerably reduces the amount and bulk of apparatus required for cable testing, office balance tests, terminal balance tests, etc. The swept-band technique provides a number of significant advantages, inasmuch as the effect of adjustments of network elements, line building-out capacitors, net building-out capacitors, and other circuit elements is immediately apparent throughout the frequency range, and the risk of improving return loss at a single frequency while degrading it at other frequencies is effectively removed.

Measurements of return loss utilizing an external test or trunk hybrid are also possible where this type of measurement must be made. In this case, the instrument is used as a conventional transmission measuring set with its output connected to the input jack of the hybrid and its input connected to the output jack of the hybrid. The impedance of the transmitting and receiving sections is adjusted to correspond to the 4-wire impedance of the hybrid, and the measurement is carried out in the normal manner. In this case the trans-hybrid loss of the external hybrid must be compensated for in arriving at the value of the measured return loss.

The return-loss measuring circuit together with the impedance-measuring circuit permits the phase angle and sense, as well as the magnitude of an impedance at any given frequency, to be determined in the following manner. The unknown impedance is connected to the jack X, the sweep drive is switched off, and the transmitting section is adjusted to the frequency of interest. The magnitude of the impedance is determined by the impedance-measuring circuit. An impedance of equal magnitude is then established at the jack X by means of a decade resistance. The instrument is then switched to the return-loss measuring circuit, the unknown is connected to the X jack, the decade resistance to the N jack and the value of the return loss is recorded. The phase angle is determined from a curve or table of phase versus return loss. After the phase angle has been ascertained, it is necessary to determine the sense of the angle, that is whether the unknown impedance contains positive or negative reactance. A reference condition is established by shunting a small capacitance across the X jack and then across the decade resistance connected to the N jack and noting the change in transmission through the hybrid coil. A decrease in return loss when the capacitance is moved from the X to the N jack indicates a positive reactance, and an increase in return loss indicates a negative reactance.

Impedance Comparison

The instrument incorporates an impedance-comparator circuit which alternately switches the two 2-terminal networks being tested into the circuit on the go and return sweeps of the transmitting section. Inasmuch as the impedance measurement function displays only the magnitude of the impedance under test, it would serve no useful purpose to measure alternately the magnitude of the unknown and standard impedances. The circuit has been modified to display the impedance of the unknown on the forward sweep and half the phasor sum of the unknown and standard on the return

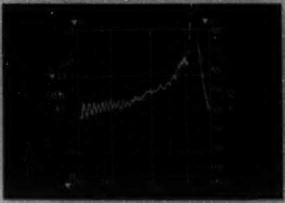


Fig. 6. Display of impedance curve of loaded toll cable 48.5 miles long, Z multiplier set to X 2

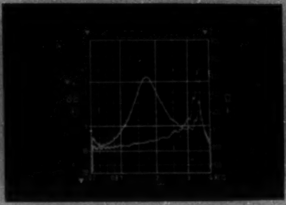


Fig. 7. Impedance comparison of line of Fig. 6 with incorrectly adjusted balancing network

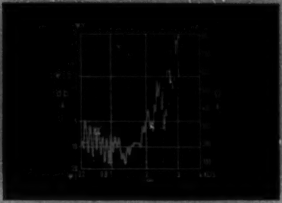


Fig. 8. Return loss of line of Fig. 6 against correctly adjusted balancing network; attenuator set to −20 decibels

sweep. The sweeps will thus coincide only at those points where the magnitude and phase of the two impedances are equal, and the measurement provides a simple indication of the relative phase of the unknown and standard impedances. This circuit offers an advantage over the return-loss circuit in that the sense of the impedance variation may be determined by observation, whereas with the return-loss measurement only the magnitude of the difference of the impedances is displayed.

LEVEL COMPARISON

Provision is included to compare the transmission loss or gain of two 4-terminal networks. For this purpose, the instrument incorporates transfer contacts on the axle of the variable capacitor which switch the high impedance input of the receiving section alternately to the *X* and *N* jacks. When this measuring facility is used, the low internal impedance of the transmitting section must be built out externally to equal the input impedance of each of the networks under test, and the networks must be terminated externally in their correct output impedance. The inputs of the specimens are thus connected in parallel through a building-out network to the output of the transmitting section, and their outputs are connected alternately to the *X* and *N* jacks on the go and return sweeps of the transmitting section. The response of the specimen connected to the *X* jack is traced out on the forward sweep and the response of the specimen connected to the *N* jack on the return sweep. The two curves may be readily compared because of the long persistence of the CRT. This technique of measurement lends itself readily to the comparison of transmission losses and to the alignment of filters, equalizers, and other networks against a standard. The accuracy of this method of measurement is limited only by the operator's ability to distinguish between the traces, which in turn is limited by the thickness of the trace. The level-comparison circuit may be selected with the function switch in the same manner as the other measuring circuits.

Applications

The level tracer is applicable to almost every class of voice-frequency measurement normally encountered in transmission engineering. A detailed consideration of the application of the instrument to cable acceptance testing is included to illustrate the saving in time and the more extensive information revealed by the use of this instrument. Other applications include measurements of office cabling balance, terminal balance and subscriber loops; the installation and adjustment of negative-impedance and hybrid type repeaters; singing-point measurements and measurements of individual message channels of multichannel transmission systems; comparison of insertion losses, filter responses and transmission characteristics; measurements of structural return loss, adjustment of balancing networks, and adjustment of equalizers. The instrument is of particular assistance in carrying out the majority of measurements required for facilities selected for data transmission.

CABLE ACCEPTANCE TESTING

The acceptance testing of toll and exchange cables normally consists of these steps following the installation of the cable: conductor-resistance measurement, return-loss measurement, and attenuation measurement. The resistance measurement is normally performed with a Wheatstone bridge and is outside the scope of this discussion. The return-loss measurement is variously performed using a noise generator and a noise measuring set, a swept-band generator with integrating detector, or at single frequencies, using standard transmission measuring equipment and a test hybrid. The attenuation measurements are normally made at frequencies of 1 kc and 60 kc. Impedance irregularity tests are usually made only if the tally of the measured return losses indicates the presence of loading faults in certain cable pairs. The application of the level tracer to this class of measurements provides a significant improvement in both the quality and quantity of information yielded and a major reduction in measurement time.

A test connection is first established as illustrated in Fig. 5. It is assumed that the distant end section of the facility is accurately known and that it has been correctly terminated and built out. The level tracer is switched to the impedance-measuring circuit, and the impedance versus frequency characteristic of the cable pair is displayed as in Fig. 6 which shows a loaded toll cable 48.5 miles long, with the multiplier set to ×2. The absence of a correct termination or the presence of line faults is immediately revealed by the shape of the measured curve. The level tracer is switched to the impedance comparison circuit and the impedance of the line is compared with the impedance of the balancing network, Fig. 7. The balancing network is then adjusted for optimum trace coincidence. The instrument is then switched to the return loss measuring circuit, and the return loss of the facility is displayed as a function of frequency (Fig. 8). Final adjustment of the variable elements of the balancing network or the line building-out elements may be made to achieve optimum return loss. For serial measurements of pairs in the same cable, it will only be necessary to transfer the far end termination and to switch the instrument to the impedance-measuring circuit and return-loss measuring circuit for each pair measured. A tolerance scheme for minimum acceptable return loss may be readily marked on the graticle. Attenuation measurements may be made by the use of two level tracers back to back for the measurement of two cable pairs at one time, or by the use of one level tracer

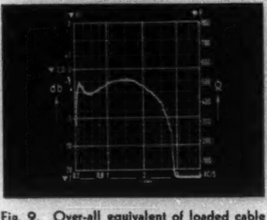


Fig. 9. Over-all equivalent of loaded cable pair (H 175)

and an auxiliary swept-band generator for the measurement of one cable pair at a time. An attenuation versus frequency curve of a loaded cable pair is shown in Fig. 9. The 60-kc transmission loss measurement for nonloaded pairs may generally be omitted as impedance measurement immediately reveals the presence of a loading coil in a nonloaded pair, and any deviation from the characteristic impedance curve of a nonloaded facility provides an indication that further testing is required. The measurement results may either be recorded photographically by means of a miniature camera or a Polaroid Land camera, or traced directly on calibrated transparencies inserted in lieu of the graticle. The advantages and the results yielded by the application of this technique of measurement to cable acceptance testing are readily apparent.

Conclusions

This paper summarizes the development problems and solutions, and the functions, and applications of a general purpose visual voice-frequency transmission-measuring set. The instrument represents an attempt on the part of its designers to reduce significantly the measurement time required for most classes of voice-frequency communications measurements, while at the same time yielding significantly greater quantities of information. Additional measuring sets utilizing the same techniques and providing similar measuring functions have been developed for the frequency ranges 30 cps–20 kc, 200 cps–6 kc, 300 cps–1,220 kc, and 10 kc–17 megacycles.

References

1. The Level Tracer, a New Measuring Device for Communication Engineering, **Herbert Knapp, Erich Alsleben.** *Siemens Review*, Erlangen, Germany, vol. 21, Jan. 1954, pp. 8–15.

2. Die Dynamischen Uebertragungsgroessen eines Vierpols bei Frequenzmodulation, **G. Bosse.** *Archiv der Elektrischen Übertragung*, Stuttgart, Germany, vol. 5, May 1951, p. 237.

3. Ueber neuere Fehlermessverfahren an pupinisierten Leitungen, **K. Kuepfmueller.** *Telegraphen- und Fernsprech- Technik*, Berlin, Germany, vol. 11, June 1922, p. 45.

A Stitched-Wiring Process for Miniaturized Communications Electronics

JOHN M. COFFIN
NONMEMBER AIEE

L. B. JOHNSON
MEMBER AIEE

WITH THE ADVENT of transistors and similar solid-state components, an investigation was undertaken at Lenkurt to determine whether printed or etched circuit boards were practical for telephone-type multichannel carrier equipment. This study resulted in development of the manufacturing process here described. It permits automatic or semiautomatic fabrication, maintains excellent standards of reliability in the finished product, and offers a degree of repairability comparable to more conventional wiring.

The stitched-wiring process, as it is called, has proved to be more economical for Lenkurt than fabrication by other methods. Although no attempt had been made to meet MIL (military) specifications with these stitched circuit boards, individual pieces and processes—for instance, the gold-plating on connector pins—are up to MIL standards. Fig. 1 shows typical circuit boards, fabricated by the stitched-wiring technique. Large components may be accommodated by spacing two or more boards at some distance apart on a modular basis.

The Process

An insulating board is used, as in printed or etched wiring. The basic board is black Grade *XP* phenolic laminate—the same material used for conventional hand-wired boards. It is manufactured by usual sheet metal shop methods. Boards are sheared or blanked to size. Common holes are punched or drilled, and, at this point, the boards may be stocked as basic inventory for stitched wiring.

The process dispenses with laminated copper foil for the wiring, and substitutes tin-coated bronze staples and automatic hand-wiring at appropriate locations. Each staple becomes the termination for one or more nearby components, and for bare-wire connections to other such terminations. As in printed wiring, all components are located on one side of the board, and wiring and interconnections on the other. Component leads and connecting wire are wrapped firmly around the staple legs, thus providing excellent mechanical and electrical connections *before* soldering. Fig. 2 is a schematic representation of the principle. Unlike conventional printed circuit assemblies, the solder is not used to strengthen the mechanical joint between components and connecting wire. Staples are driven through the board without prepunching. Component leads pass through adjacent holes and wrap around staple posts. Staples are interconnected with bare-wire straps and the entire underside of the board is dip soldered.

As many components as possible are selected or designed to have stiff leads or posts, instead of flexible leads. For instance, a transformer may have five stiff posts projecting from the bottom of the case, similar to the pins of a vacuum tube. These posts serve as terminals and also provide means for attaching the transformer to the board. All components are mounted on the board and wired complete, entirely independent of the metal mechanics.

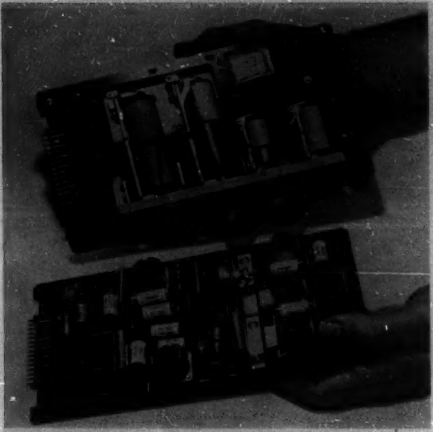
The stitched-wiring process begins with the attachment of male connector pins to the edge of the board (if the board is a plug-in unit). These connector pins are preforms of two different sizes which,

---

Paper 61-1046, recommended by the AIEE Wire Communication Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted June 14, 1961; made available for printing August 8, 1961.

John M. Coffin and L. B. Johnson are with Lenkurt Electric Company, Inc., San Carlos, Calif.

The authors gratefully acknowledge the valuable assistance of several Lenkurt engineers, who contributed to the development of the stitched-wiring process and its reduction to practice. Among these are: A. G. Ewing, John R. Feineman, G. B. Hanford, O. A. Keyer, John G. Lord, W. C. Matthews, J. R. Morison, John E. Ryan, and David N. Winslett.



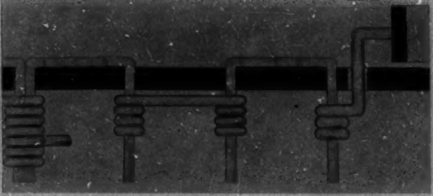


Fig. 1 (left). Typical stitched-wiring circuit boards

Fig. 2 (above). Stitched-wiring principle illustrated

when crimped and locked into place, form an even row of contacts. The insertable portion of the connector pin is closely controlled to 1/16-inch thickness to meet the same specifications as commercial prefabricated male connectors. Within the edge and perpendicular to the face of the board, a double row of connector posts is formed, standing about 3/16-inch high. A section of such a connector is shown in Fig. 3. After insertion in the board, connector pins are treated in the same manner as component mounting staples. Gold-plating on the pins insures maintenance of good contact under adverse conditions.

The next operation involves driving the tin-coated bronze staples at precise, discrete positions in the board, thus creating staple posts. The location of each such post represents the termination of a component lead, a bare-wire strap or, infrequently, a long, insulated crossover, or jumper. All staples pass through the board in the same direction, with the crowns appearing on the component side of the board. Fig. 4 shows part of the staple-driving machine, which dispenses with the need for prepunching. An anvil beneath retracts downward to permit moving the board horizontally to the next location.

Just as with the connector posts, the staple legs are not clinched, but project through the board about 3/16 of an inch. Holes, through which one or more component leads may be passed, are sufficiently close to each staple post so that Keller or similar automatic wire-wrapping guns can wrap the component leads around the adjacent posts.

With staples and connectors added, the appearance is roughly that of a printed circuit board which has been etched and drilled, but is still without components. In the stitched-wiring process, components may be inserted manually or, if using standard component insertion heads as in Fig. 5, automatically. Two special pieces of equipment are used at component insertion stations:

Fig. 3. Close-up of preformed male connector pins

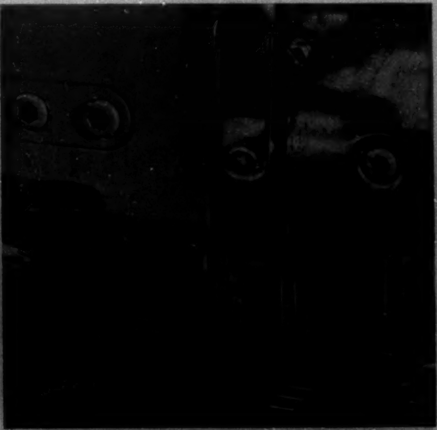


Fig. 4. Close-up of stitching staple driving head

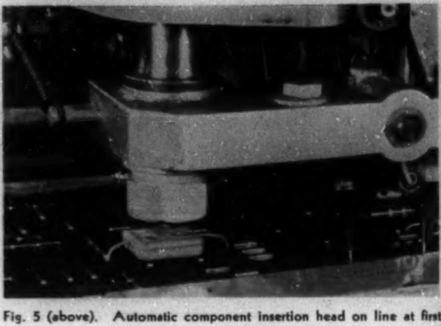


Fig. 5 (above). Automatic component insertion head on line at first station, showing resistor half inserted

Fig. 6 (right). Wrapping head, wrapping the leads of a component around adjacent staples

1. An automatic head containing two wire-wrapping bits of the end-loading-type, which automatically rise into position and simultaneously wrap the two component leads around adjacent staple posts. Fig. 6 shows a wrapping head in position beneath a stapled board. The same head is used at the manual stations, and for the full range of component sizes. The bits retract after wrapping to permit the conveyor to advance.

2. An automatic head containing two side-loading wire-wrapping bits and a spool of bare wire. The bits rise into position, when the conveyor is in dwell, and automatically strap a piece of bare wire across any two staple or connector posts; then cut the spool wire free. The heads reload themselves with wire for the next board during the time the conveyor is in motion. Fig. 7 shows a typical strapping head. Strapping bits rise, carrying a length of bare wire, and wrap both ends around two staple posts to be strapped together. Rotation is equal and opposite. The cycle is fully automatic.

Both of these heads are self-sequencing, but are given a starting signal by the conveyor itself. When the board is in position, both types of head rise to the bottom of the board so that the staple or connector posts enter the center holes of the wrapping or strapping bit. When the operation is completed, a matter of approximately half a second, the bits retract so that the conveyor can advance to the next station. The strapping stations are unattended, and the only equipment installed above the conveyor line at these points is a pressure pad, shown in Fig. 8, which moves into place and holds the board flat as the heads are operating. An automatic set of holding fingers can be adjusted to fit any board and any length strap so that two steel fingers are directly above strapping bits.

A stitched-wiring assembly line, then, not only serves as a component insertion line, as it would with printed circuits,

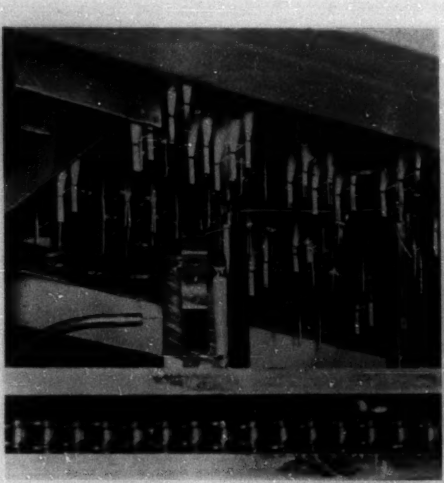


Fig. 7 (left). Typical strapping head

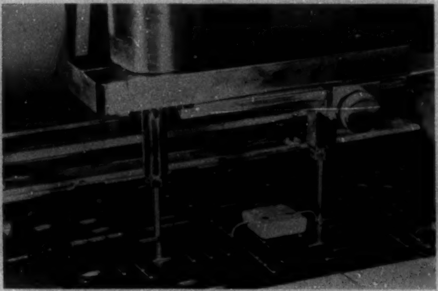


Fig. 8 (below). Automatic fingers which hold board during strapping but also to provide necessary interconnections between components.

The basic stitched-wiring machinery, shown in Fig. 9, is a General Mills Autofab conveyor line, originally designed for automatic insertion of components in printed circuit boards, but modified in its arrangement as subsequently described:

1. An operator may sit at any position along the conveyor with comfort.
2. Wrapping and strapping stations installed under the conveyor are as easily set up and put into proper positions as the insertion heads themselves. A pancake design provides adequate clearance above an operator's knees.
3. The frame of an insertion head can be used to hold and actuate the pressure pad required at the strapping stations.
4. Automatic insertion is attained with the same heads that would be used for printed circuits, except that a slight modification is required to provide a longer lead length. The wrapping stations replace the anvils originally used to clinch the leads on a printed circuit board.
5. Setup is sufficiently fast that multiple-pass operations are economical.

The soldering station is the same type as would be used with a printed circuit board. The bottom of the board exposes only staple posts, bare-wire straps, and ends of the component leads. Passing the board through a soldering station adequately solders all connections.

Unlike printed circuits, these boards can be tested before soldering, since good mechanical and electrical connection of all wire and leads is provided. Test fixtures can be made which use the staple posts as connectors.

In passing the board through the soldering station, no great care is necessary to avoid jiggling or vibration, since each mechanical connection not only is very tight but provides almost perfect dimensions for capillary attraction of the solder into the spaces between the wrapped wire and the posts themselves.

Automatic and semiautomatic machinery is not required to manufacture these boards—they can be completely built by hand. Field repair is simple, and nothing more than ordinary hand tools and a soldering iron are required to replace any defective component on a board.

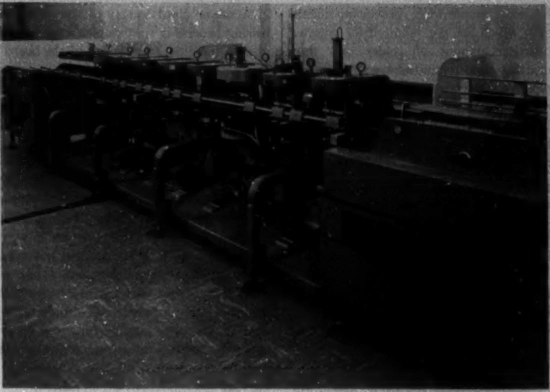


Fig. 9. General view of combination strapping and component insertion line

Features and Advantages of Stitched Wiring

Stitched wiring is surveyed on the basis of mechanical, electrical, and economic considerations, and these facts emerge:

1. Stitched-wiring base material is inexpensive.
   a. *XP* or *PC* phenolic laminate, as opposed to *XXXP*.
   b. Stitched-wiring material is not copper-faced.
2. Holes for component leads are punched in plain base material without the problems that prevail when a composite material is used.
3. A stitched-wiring board is ready to assemble when the holes have been punched and connectors and staples driven. It may be assembled:
   a. Automatically.
   b. Semiautomatically.
   c. By hand.
4. The board may be tested completely before soldering. For automatic testing, the staple posts act as connector pins and may be plugged into a test jig.
5. Automatic soldering is the same in any case, but with the stitched-wiring method there is the least copper dilution in the solder pot.
6. For short runs, the start-up cost is low—there is only the variable of the discrete positions for the staples. Two templates are required for each board at about $55 each. Without a programmable drilling machine, a drill jig will be needed.

In terms of capital for equipment, a minimum investment is required for these items:

1. Standard drill press.
2. Drill jig for each board, constructed to permit drilling eight to ten boards at one time; cost, approximately $160.
3. Two stapling templates for each board; total cost, approximately $110.
4. Stapling machine (which could be simpler for very short runs); cost, as itemized.

Wales duplicator attachment	$1,500
Special mounting base	4,500
Acme stitching head	1,000
Assembly and alignment labor	1,000
Total	$8,000

From this point on, additional equipment accomplishes cost reduction only. Lenkurt of Canada has built a simpler machine for short runs which might cost $3,000 in the United States. Substantially, the listed items are nearly identical for all processes used in manufacturing circuit boards: component preparation and insertion, automatic fluxing and soldering, hole drilling or punching, and designation marking.

Cheaper for the stitched board than for any other method are: base material,

Table I. Estimated Factory Cost of Stitched-Wiring Boards (Small to Medium Quantities)

Material	5- by 8-Inch Board	$0.1240
	8-Pin Connector	0.0872
	Staples, Wire (67)	0.1340
	Guide Clips (2)	0.0156
	Total Material Cost	$0.3608
Labor, Hours	Shear	0.0195
	Punch	0.0040
		0.0072
		0.0016
	Silk Screen, Designations	0.0110
	Neutralize, Wash	0.0024
	Drill (132 versus 76)	0.0528
	Stitch (67)	0.0536
	Install Connector (8)	0.0216
	Clean	0.0024
	Guide Clips	0.0117
	Bare-Wire Straps (12)	0.0071
	Total Labor Hours	0.1949
	Labor Dollars at $2.00 per hour	$0.3898
	Burden Dollars at 165%	0.6432
	Manufacturing Cost (Material, Labor, and Burden)	1.394
	Reject and Rework Cost at 1%	0.014
Cost of Board Ready for Components		$1.408

From here on, cost of component insertion and dip soldering can be computed more accurately by each individual on the basis of what components he uses and whether or not he already has a soldering machine.

including staple wire; processing chemicals, which are not required for stitched boards; amount of solder and flux; drafting time for initial layout; time to repair a defect (rework); and in-process reject rate.

In terms of board Q (quality), a stitched board is superior mechanically, and it possesses high electrical Q in that:

1. Conductors are not in intimate contact with the board; therefore, moisture absorption had almost no effect on the Q of the board.
2. Distributed capacitance is low.

From a board design point of view:

1. Crossovers are easily attained, yielding the advantages of a double-sided etched board.
2. Strapping options are easily attained.
3. Slightly dissimilar boards require no extra tools—options are inherently available in the drill jigs and templates.
4. A circuit design change makes nothing obsolete; the tools and setups are easily modified—all of which gives board designers much freedom.

A minor disadvantage of the stitched-wiring process is that staples may not be driven too close to the edge of the board.

Cost Reduction

Some steps, taken to reduce costs, produce a much higher economic return than others. For instance, automatic soldering provides the highest return of all; automatic component insertion the lowest. The combination of manual insertion on the conveyor line and automatic wire wrap has a high return and, in this area of part automation and part manual operation, the greatest earnings are with small and medium quantities. Complete automation is quite feasible for very large quantities with commensurate cost reduction.

Cost Comparison

There is no satisfactory manner for comparing the cost of a stitched-wiring with a printed circuit board. Finished boards with all components assembled and soldered and the circuit tested should be examined as finished products, but quantities, economical lot sizes, labor rates, product mix, capital investment, and original development costs would become involved in calculations beyond the scope of this paper. Therefore, our cash output is presented as the best estimate of average cost now available in the United States for producing stitched-wiring circuit boards in small to medium quantities or batches. See Table I.

Conclusions

The stitched-wiring technique is now being used in the manufacture of multichannel telephone and telegraph carrier equipment, and various supporting types of communication equipment. Substantial manufacturing economies have been realized, and long-term observation of this equipment confirms that reliability and serviceability are excellent.

◆

Discussion

R. W. Perkinpine (American Telephone & Telegraph Company, Kansas City, Mo.): I have noted with interest the paper concerning a stitched-wiring process by John M. Coffin and L. B. Johnson.

Their plan provides a novel and most interesting approach to the problem of providing solid wiring and soldered connections. One advantage over printed circuit arrangements is that there would be no foil to loosen up on the soldering side of the board when components had to be replaced. One objection to the stitched-wiring process might be that it requires considerable space on both sides of the board. In locations where space is at a premium, this might present a problem.

It appears to me that this stitched-wiring approach has some advantages also over the boards which are equipped with terminals for the components. Although such boards permit of good soldering, they are not as adaptable to multiple usage as the punched boards.

John M. Coffin and **L. B. Johnson**: We agree that foil loosening or delamination, a disadvantage of printed or etched circuits, is not encountered in the stitched-wiring technique. In fact, the mechanical stability of a stitched-wiring board is so good that the base material can be considered the weakest part of the structure. Delamination of printed boards does not occur only when replacing components, but has occurred with the base material immediately upon shearing. This is because of poor quality adhesive, sometimes over a large area.

Stitched boards, of course, use no adhesive, and the excellent mechanical strength and stiffness of the staples permit changing any component any number of times. In an extreme case, a staple itself can be driven out backwards and be replaced before putting new components back on the board. Circuit changes, then, are feasible too, remembering that the circuitry is completed with simple bare-wire straps which can be changed or re-routed.

The space problem is not as severe as it might appear—component density on stitched boards produced at Lenkurt compares favorably with that of etched circuits for the military, also produced here. However, along the Z dimension—component height plus board thickness—the stitched board does require additional height for the staple legs. If space is of extreme importance, staple legs can be shortened, at some expense of assembly labor, to 1/8 of an inch. A few of our early prototypes used this leg length.

Our older systems, some of them designed more than 10 years ago, use the terminal-lug technique referred to as affording comparable solderability. We realize that such boards have limited usage, and hasten to add that they are wasteful of space and unnecessarily costly. The multiple usage of the stitched boards has exceeded our expectations, and great credit is due our designers who, required to use stitched wiring, solved complex space assignment problems without violating the ground rules. W. W. Frank did an excellent job on the diode matrix panel in our exchange carrier system—180 diodes appear in a treble panel, each card of which is our standard size of approximately 5 by 9 inches including handle and connector, with all the diodes mounted on two of the cards, the third card providing a male connector on a blank card which acts as a mechanical shield.

We appreciate having had the opportunity to expand our subject, and shall be happy to answer any direct inquiries from those who may wish more details about the process.

Electrical and Pressure Losses in a Magnetohydrodynamic Channel Due to End Current Loops

GEORGE W. SUTTON
NONMEMBER AIEE

H. HURWITZ, JR.
NONMEMBER AIEE

H. PORITSKY
MEMBER AIEE

Summary: End electrical and pressure losses are calculated for d-c magnetohydrodynamic (MHD) power generators, accelerators, or induction pumps having a geometry of a rectangular parallelopiped with the electric field, magnetic field, and fluid flow mutually perpendicular. For an MHD generator, having a cubic geometry, the maximum conversion efficiency is only 29%. This can be improved by increasing the length of the generator, inserting vanes, and extending the magnetic field, so that efficiencies of 90% should be attainable.

THE END LOSSES in a conducting fluid have been studied for uniform flow in a channel of rectangular cross section with transverse magnetic fields and electric fields of finite extent such as occur in d-c induction pumps or MHD generators. These end losses are caused by the gradients in the magnetic field at the inlet and outlet which cause eddy currents in the fluid. The required expressions for the electric field distribution are obtained by means of conformal mapping, from which the electric currents, electric power, and pressure drop are determined. Finally, pump and generator efficiencies are calculated.

Three cases are considered:

Magnetic field termination at the end of the electrodes

Constant magnetic field extension beyond the ends of the electrodes

Exponentially decreasing extension of the magnetic field beyond the ends of the electrodes

It is found that the end losses decrease with either increasing aspect ratio (ratio of electrode length to channel width) or decreasing electric current. Magnetic field extensions decrease the losses, the exponential extension being superior to constant magnitude extensions. Insulating vanes placed in the flow upsteam and downstream of the electrodes will also decrease the losses by increasing the apparent aspect ratio.

Nomenclature

a=channel aspect ratio$=2Y_0/h=y_0/\pi$
B=magnetic field
c=dimensionless coordinate of end of constant magnitude magnetic field extension
C_p=specific heat
E=electric field
h=distance between electrodes (channel width)
J=total electric current
$\mathbf{j}$=electric current vector
k=argument of elliptic integral
$k'=(1-k^2)^{1/2}$
K=complete elliptic integral of first kind, $K(l, k)$
K'=complete elliptic integral of first kind, $K'(l, k')$
p=pressure
P_i=ideal generated or consumed electric power
$P_a{}^+$=actual power generated or consumed
P_L=power loss due to termination of magnetic field at ends of electrodes
ΔP^+=additional power generated by extension of magnetic field
$\mathbf{q}$=fluid flow vector
T=temperature
U=fluid velocity
V=voltage
X, Y=physical co-ordinates
x, y=nondimensional co-ordinates, $2\pi X/h$, $2\pi Y/h$
x', y'=transformed complex co-ordinates
y_e=dimensionless e-folding length of exponential magnetic field extension$=2\pi Y_e/h$
y_0=half length of electrode, dimensionless, $2\pi Y_0/h$
Y_0=half length of electrode
Z=direction of magnetic field
z, z'=complex variables
ϵ_g=generator efficiency, $P_a{}^+/\Delta p_a Uh$
ϵ_p=pump efficiency, $\Delta p_a Uh/P_a{}^+$
γ=exponential coefficient for exponential magnetic field extension$=(2y_e)^{-1}$
η=voltage coefficient
ϕ=electrical field potential
ψ=electric current flux function
ψ_L=electric current flux loss by termination of magnetic field at ends of electrodes
$\Delta\psi^+$=electric current flux added by extension of magnetic field
σ=electrical conductivity
ρ=fluid mass density
β_1—see equations 19 and 66
β_2—see equations 23 and 76
β_3—see equation 25

Subscripts

a=actual
i=ideal
oc=open circuit

Introduction

In a d-c MHD generator or induction pump, an electrically conducting fluid is passed through a rectangular channel in which a magnetic field is placed at right angles to the flow direction, and electrodes are placed on the two sides of the channel that are parallel to the magnetic field. For a pumping device, the electrodes are connected to a voltage supply, and current flows through the fluid; the interaction of the current and the magnetic field causes a Lorentz force which increases the pressure. In the generator, the electrodes are connected to an external load, and the moving conducting fluid interacts with the magnetic field to cause an induced current to flow through the external load.

Such devices usually have low efficiencies that are caused by two effects. The first loss is caused by the conducting fluid in the regions upstream and downstream of the magnetic field which acts as an electrical shunt in parallel with the fluid that is within the magnetic field. These shunting losses change the electrical characteristics of the device: for a pump more electric power is required, while for a generator less power is generated. In addition, these currents change the pressure rise or pressure drop through the device. Finally, the difference in pressure work and electric power appears as ohmic heating of the fluid, which is usually undesirable. The shunting effect usually increases this ohmic heating.

In the present paper, expressions are derived for losses, and calculations are presented for several different cases. To make the problem tractable, a number of assumptions are made which will be described later.

The other loss is due to the magnetic field which is generated by the electric current in the fluid. The effect is to decrease the magnetic field at the inlet and increase the magnetic field at the exit. The resulting nonuniformity of the magnetic field aggravates the shunting effect. However, this problem is be-

---

Paper 61-1078, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted January 16, 1961; made available for printing September 7, 1961.

GEORGE W. SUTTON is with the General Electric Company, Philadelphia, Pa., and H. HURWITZ, JR. and H. PORITSKY are both with the General Electric Company, Schenectady, N. Y.

The authors wish to thank Dr. E. Luebke for his suggestions and Miss Cathy Feucht and Mr. Sheldon Biecher for calculating the various integrals and series on the IBM (International Business Machines) *704* and LGP (Royal McBee Corporation), *30* digital computers, respectively.

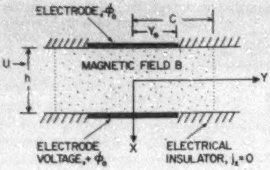


Fig. 1. Schematic diagram of magnetohydrodynamic channel

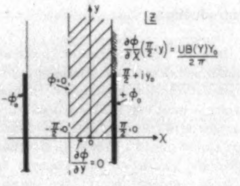

y
z
$\frac{\partial\phi}{\partial x}(\frac{\pi}{2},y) = \frac{UB(Y)Y_0}{2\pi}$
$\frac{\pi}{2}+iy_o$
$+\phi_o$
$\phi=0$
$-\phi_o$
$-\frac{\pi}{2}$,0
$\frac{\pi}{2}$,0
0
X
$\frac{\partial\phi}{\partial y}=0$

Fig. 2. Z plane showing region to be mapped

yond the scope of this paper; thus, it will be assumed in the present paper that the magnetic field that is induced by the electrode currents may be neglected. Therefore, the analysis is limited to cases where the magnetic Reynolds number, defined by $\mu_o\sigma Uh$, is small compared with unity. This condition is fulfilled in most devices in which the conducting fluid is an ionized gas because the electrical conductivity is small. However, this condition may not be fulfilled in a device in which the conducting fluid is a liquid metal.

Mathematical Model

The channel is assumed to be straight with a constant rectangular cross section. On two sides of the channel are electrodes of length $2Y_o$ and a height equal to the channel height; see Fig. 1. The voltage on one electrode is ϕ_o; on the other it is $-\phi_o$, where

$$\phi_o = \frac{1}{2}\eta UB_oh \qquad (1)$$

where η is a dimensionless loading parameter of the device. For an ideal device in which end losses do not exist, for $0<\eta<1$, the device acts as a generator, while for $\eta>1$ it is a pump. All wall surfaces except the electrodes are assumed to be perfect insulators. The electrodes are assumed to be perfect electrical conductors.

The channel width, which is the lateral distance between electrodes, is h; thus the aspect ratio a of the channel is $2Y_o/h$. Within the region bounded by the electrodes, the magnetic field is taken constant, equal to B_o. In the region upstream and downstream of the constant magnetic field region, the magnetic field varies only with the axial direction, Y. This condition is fulfilled when the channel height in the direction of the magnetic field is small, and when the pole faces of the magnet are rectangular. One designs the magnetic air gap to be as small as practical, hence this is a reasonable assumption.

The fluid density is taken constant, and the fluid velocity is taken constant, equal to U, and only in the downstream direction. Thus, the effects of viscosity are neglected. This assumption is valid when the flow is turbulent since the velocity profile is full. The electrical conductivity of the fluid is assumed to be a constant scalar.

With these assumptions, the variables do not vary in the direction of the magnetic field (Z-direction). Thus, only a region of unit height in the direction of the magnetic field is analyzed. The electrodes are at

$$X = \pm\frac{1}{2}h,\ |Y| < |Y_o|$$

The performance of an ideal device with zero end losses will be analyzed first to provide a basis for comparison. Next, the electrical losses will be calculated for a magnetic field which terminates at the end of the electrodes ($\pm Y_o$). Two extensions to the magnetic field are considered: one which has a constant magnitude, and then decreases to zero at $Y=C$; and one which decreases exponentially with an e-folding length Y_e.

Expressions for the pressure drop are derived; these are used together with the electrical characteristics to calculate the efficiencies for both pumps and generators. Finally, expressions are given for the ohmic heating.

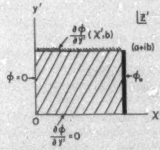

y'
z'
$\frac{\partial\phi}{\partial y'}(X',b)$
(a+ib)
$\phi=0$
ϕ_o
0
X'
$\frac{\partial\phi}{\partial y'}=0$

Fig. 3 (above). Z′ plane

$Z'=(a/K)\,\mathrm{sn}^{-1}(w,k)$
$w=(1/2)\sin(Z+1)$
$k=2^{1/2}(1+\cosh y_0)^{-1/2}$
$b/a=K'/K$

The vector current equation is taken as

$$\mathbf{j}=\sigma(\mathbf{E}+\mathbf{Q}\times\mathbf{B}) \qquad (2)$$

Since $\mathbf{B}$ is steady, $\nabla\times\mathbf{E}=0$ and $\mathbf{E}$ may be expressed in terms of an electric potential: $\mathbf{E}=-\nabla\phi$. Upon taking the divergence of equation 2 one obtains

$$\nabla^2\phi=0 \qquad (3)$$

The value of ϕ is specified on the electrodes; thus

$|Y|<Y_o$:

$$\phi\left(\pm\frac{1}{2}h,\ Y\right)=\pm\phi_o=\pm\frac{1}{2}\eta UB_oh \qquad (4)$$

Upstream and downstream of the electrodes, the current to the walls at $X=\pm 1/2h$ is zero; from equation 2 this becomes

$$|Y|>Y_o,\qquad X=\pm\frac{1}{2}h:$$

$$\partial\phi/\partial X=UB(Y) \qquad (5)$$

Thus, equation 3 must be solved subject to the boundary conditions of equations 4 and 5.

The electric power is given by the product of the total current and voltage difference; the pressure drop is given by the total Lorentz force J_zB; and the ohmic heating is the difference between the pressure-work and electric power.

Conformal mapping of the X-Y plane is used to solve this problem; see Figs. 2 and 3. Only the results are presented in the main text of the paper; the details are contained in appendixes. In addition, the following dimensional coordinates are used:

$$x=2\pi X/h$$

$$y=2\pi Y/h$$

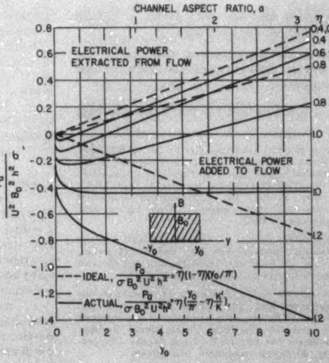


Fig. 4 (right). Electric power, magnetic field ends at electrodes

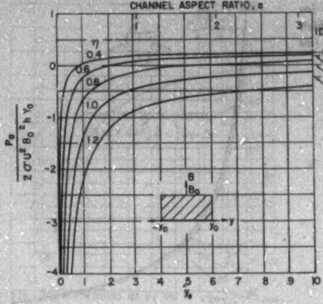


Fig. 5 (left). Electric power density constant, magnetic field ends at electrodes

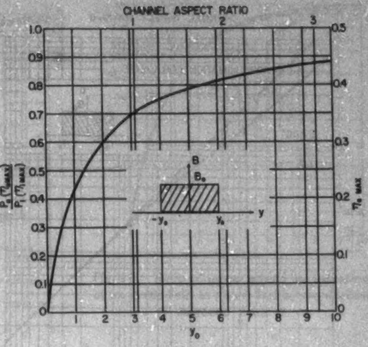


Fig. 6 (right). Ratio of actual maximum power density to ideal maximum power, and generator coefficient

Constant magnetic field ends at electrodes

The channel width is 2π while the channel length is $2y_0$; thus, the aspect ratio is $a = y_0/\pi$.

Ideal Generator or Pump

The ideal generator is defined as one in which there are no end effects; it is useful to obtain relations for the desired quantities for the ideal generator for later reference. The electric field in the fluid in an ideal generator is given by the sum of the static electric field and the magnetically induced electric field. Both are in the X-direction only, thus conformal mapping is not required. The current density in the X-direction is the product of this field and the electrical conductivity, or

$$j_x = \sigma(E_x + UB_0) \tag{6}$$

In the ideal generator the components of the current density in the other directions are zero. In accordance with equation 1, the voltage across the two electrodes is given by

$$V = \phi_0 - (-\phi_0) = \eta UB_0 h \tag{7}$$

so that the static electric field associated with the electrodes is

$$E_x = -\eta UB_0 \tag{8}$$

Substitution of equation 8 into equation 6 yields

$$j_x = (1-\eta)\sigma UB_0 \tag{9}$$

The electric power produced per unit volume is $-\mathbf{E}\cdot\mathbf{j}$ which in this case is $-E_x j_x$. The total power per unit height is obtained by multiplying the power density by the area of the plan-form. From equations 8 and 9 the ideal power produced is

$$P_i = \eta(1-\eta)\sigma U^2 B_0^2 \cdot 2Y_0 h \tag{10}$$

The ideal pressure drop is given by the total Lorentz force on the fluid, $2Y_0 j_x B_0$, or

$$\Delta p_i = (\eta-1)\sigma UB_0^2 \cdot 2Y_0 \tag{11}$$

The efficiency of the device is given by the ratio of the flow work $Uh\Delta p_i$ and the power output. For a pump or generator, respectively, the efficiency is thus

$$(e_p)_i = \eta^{-1}$$

with

$$\eta > 0 \tag{12}$$

for the pump and

$$(e_g)_i = \eta$$

with

$$0 < \eta < 1$$

for the generator. The ohmic heating per unit volume is the difference of the flow work and electric power, or j^2/σ; the total temperature rise of the fluid as it passes through the generator is given by

$$C_v \Delta T = 2Y_0 j_x^2/\sigma\rho U = 2\sigma\rho^{-1} Y_0 (1-\eta)^2 UB_0^2 \tag{13}$$

which shows that for both generators and pumps it is desirable to maintain η as close to unity as possible; however, as η is brought close to unity, the net power decreases as is seen from equation 10; but the end losses are proportional to η. Thus the end losses do not allow operation of η close to unity. As will be shown subsequently, this is quite critical.

Electrical Characteristics

Magnetic Field Terminates at $Y = Y_0$

The general solution of the problem for any aspect ratio is given in Appendix I. However, for values of the aspect ratio greater than about 0.3, the results reduce to a simple form; the following equations are based on this. The total current is given by

$$J = 2\sigma\{UB_0 Y_0 - \phi_0[a + (2/\pi)\ln 2]\} \tag{14}$$

where $2\phi_0$ is the load voltage. The open-circuit voltage is given by $2\phi_0$ when $J=0$, or

$$V_{oc} = UB_0 h[1 + (2/a\pi)\ln 2]^{-1} \tag{15}$$

Thus the value of the nondimensional open-circuit voltage, η, is less than the ideal value which is unity.

The power is given by the product of equation 14 and the electrode voltage difference; in terms of η

$$P_a = \eta\sigma U^2 B_0^2 h^2 [a(1-\eta) - (2\eta/\pi)\ln 2] \tag{16}$$

is obtained. Equation 16 is shown in Fig. 4 as the solid lines, for various aspect ratios. The curvature for $y_0 < 1$ is due to the mutual influence of the end effects; see equation 46. Values of P_a less than unit indicate a net power consumption. Thus, if the aspect ratio of the channel is too small, the channel cannot be used as a generator. As the aspect ratio increases, the device can act as a generator for low values of η. If the channel aspect ratio is 2, notice that very little power can be extracted for a loading factor of $\eta = 0.8$.

Also in Fig. 4 the ideal electric powers are indicated by the broken lines. It may be seen that the electrical loss, which is the difference between the actual and ideal theoretical power, is proportional to η. This is to be expected because the greater the value of η, the greater the electrode potential, and hence the larger the electrical loss. Note, that for values of y_0 greater than 2, the slope of the power curve is very close to the ideal.

The average power density based on the region enclosed by the electrodes is shown in Fig. 5. Of course, power is actually dissipated in the region external to this region. For large values of y_0, the aver-

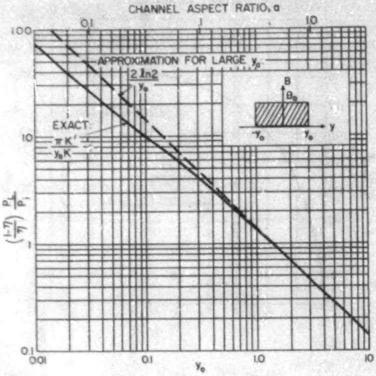


Fig. 7 (left). Power loss for constant magnetic field which terminates at the end of the electrodes

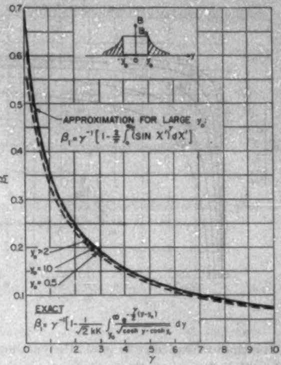


Fig. 8 (right). β_1 **versus** γ **for** $B = B_0 e^{-(1/2)(y-y_0)\gamma}$

age power density approaches the ideal, as indicated by the broken lines on the right side of Fig. 5.

For a generator, it may be desired to operate at a value of η at which the power output is a maximum; this is obtained by differentiation of equation 14. The resulting values of η and the ratio of the maximum power to the ideal maximum power are shown in Fig. 6. For channel aspect ratios as high as 10, the maximum power is still 12% less than the ideal power. If, on the other hand, one attempts to keep $\eta = 0.5$, then the power output is considerably less than maximum especially for small channel lengths. This is quite evident from Fig. 4.

Finally, it is desirable to calculate the "lost" power, which is the difference between the ideal and actual. The ratio of the "lost" power to the ideal is given by equation 50 which becomes

$$P_L/P_t = 2\eta(1-\eta)^{-1}\pi^{-1}a^{-1}\ln 2 \tag{17}$$

This is shown in Fig. 7 for various aspect ratios, as well as the exact theory given in Appendix I.

For values of η close to unity, which are desirable to minimize pressure drop, the end electrical losses given by equation 17 are excessive, except for very large aspect ratio channels ($y_0/\pi >> 1$). That is, for $\eta \sim 1$ the channel acts as an inefficient generator or as a pump, for much larger amounts of power are required to obtain a given pressure increase. Hence it is desirable to extend the magnetic field beyond the electrodes to minimize these losses. The effect of extending the magnetic field is considered next.

EXPONENTIAL MAGNETIC FIELD EXTENSION

Magnetic fields do not actually end abruptly; they decrease in a manner which is approximately exponential. The e-folding length of the decrease, Y_e, may be controlled by the design of the channel and magnetic field. Thus, in the region beyond the electrodes, the magnetic field may be taken as

$$B = B_0 \exp\{-[(|Y|-Y_0)/Y_e]\} \tag{18}$$

This extension is assumed to be both upstream and downstream of the main magnetic field. The increase in electrode current as obtained from equation 65 is

$$\Delta J^+ = -2\sigma U B_0 h \beta_1 \tag{19}$$

where for aspect ratios greater than about 0.3,

$$\beta_1 = \gamma^{-1}\left[1-\pi^{-1/2}\Gamma\left(\frac{1}{2}\gamma+\frac{1}{2}\right)/\Gamma\left(\frac{1}{2}\gamma+1\right)\right] \tag{20}$$

where $\gamma = 2/y_e$ and $y_e = 2\pi Y_e/h$. Values of β_1 are shown by the solid line in Fig. 8. The additional power, ΔP_+, is obtained

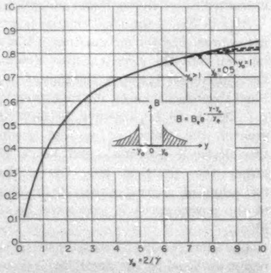


Fig. 9. Ratio of power gained by exponential extension of magnetic field to power lost by termination at end of electrode

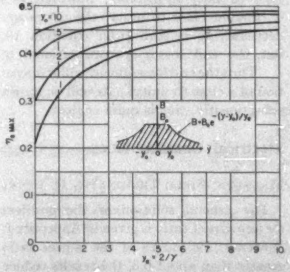


Fig. 10. Generator coefficient for maximum power density for exponential magnetic field extension

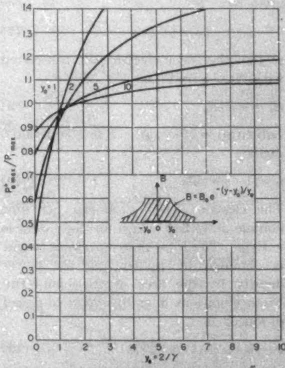


Fig. 11. Total maximum generated power density for various exponential magnetic field extensions

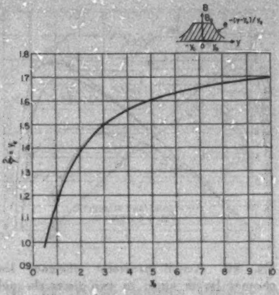


Fig. 12. Values of y_e for $P^+_{a\,\max} = P_{i\,\max}$ for exponential extension of magnetic field

by multiplying ΔJ_+ (equation 19) by the electrode potential difference, $2\phi_0$. The ratio of the additional power to the "lost" power is shown in Fig. 9. As the *e*-folding length becomes very large, the value of $\eta \Delta P_+/P_L$ approaches unity. This means that for a generator, when $\eta<1$, the additional power exceeds the power loss; but for a pump, where $\eta>1$, the additional power is always less than the power loss.

The actual total power is given by the sum of the additional power and equation 14, as follows:

$$P_a{}^+ = \sigma\eta U^2 B_0{}^2 h^2 [a(1-\eta) - \eta(2/\pi)\ln 2 + 2\beta_1/\pi] \quad (21)$$

As before, it is possible to maximize the power output of a generator with respect to η. Values of η that maximize the power are shown in Fig. 10, and the corresponding values of the maximum power are shown in Fig. 11. For sufficiently large extensions, the maximum power can exceed the ideal maximum. This effect is greatest for channels of small aspect ratio: the regions upstream and downstream of the electrodes are also acting as a generator and are feeding power to the electrodes.

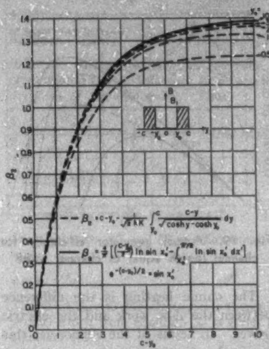


Fig. 13. β_2 versus $(c-y_0)$ for constant magnetic field extension

The values of y_e for which the actual maximum power equals the ideal maximum power are shown in Fig. 12. It is seen that the larger the aspect ratio, the less the additional extension in order to obtain the same power output as the ideal generator.

Finally, the open-circuit voltage is given by

$$V_{oc} = UB_0 h(1+2a^{-1}\pi^{-1}\beta_1)\times(1+2a^{-1}\pi^{-1}\ln 2)^{-1} \quad (22)$$

Constant Magnitude Extension of the Magnetic Field

Another method of decreasing the end loss is to use a longer magnet. For this case, it is assumed that in the region upstream and downstream of the main magnetic field the magnetic field is B_1, and extends from $Y=Y_0$ to $Y=C$. This causes an additional electrode current of

$$\Delta J^+ = -2\sigma U B_1 h \beta_2/\pi \quad (23)$$

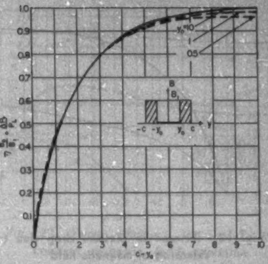


Fig. 14. Ratio of power added by extension of magnetic field to power lost by termination of magnetic field at y_0

Values of β_2 are given in Fig. 13 in terms of $Y_0 = 2\pi Y_0/h$ and $c = 2\pi C/h$. The ratio of the additional power to the "lost" power when the magnetic field terminates at the electrode ends is shown in Fig. 14. The values of η which lead to maximum power generation are shown in Fig. 15. Finally, the ratio of the total actual power to the ideal power is shown in Fig. 16. The values of $c-y_0$ corresponding to $(P_a{}^+)_{(\max)} = (P_i)_{(\max)}$ are shown in Fig. 17. These correspond, of course, only to a generator.

Pressure Change and Efficiencies

The pressure change through the channel is given by the total Lorentz force averaged over a cross section as follows:

$$\Delta p_a = h^{-1}\int_{-\infty}^{\infty}\int_{-h/2}^{h/2} j_z B\,dX\,dY \quad (24)$$

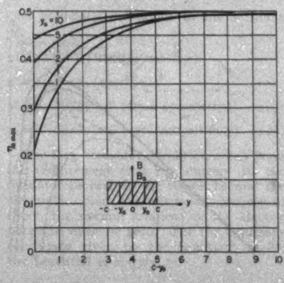


Fig. 15. Generator coefficient for maximum power density for constant magnetic field extension

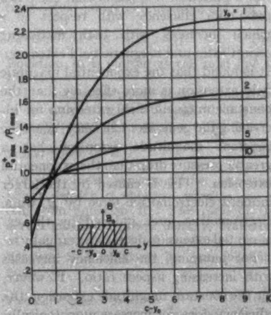


Fig. 16. Total maximum power density for various constant magnetic field extensions

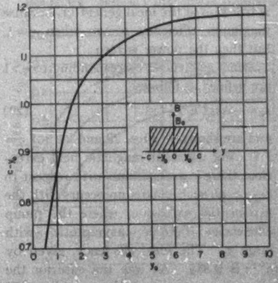


Fig. 17. Constant magnetic field extension required for actual maximum power equal to ideal maximum power

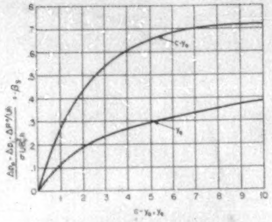


Fig. 18. Increase in pressure drop due to extension of magnetic field

The details of the calculation of Δp_a are given in Appendix II. The result is

$$\Delta p_a = \Delta P^+/Uh + \sigma U B_o^2 h\beta_3 + \Delta p_i \qquad (25)$$

The values of ΔP^+ are given by the product of $2\phi_o$ and the currents given by equations 19 or 23. The value of the ideal pressure drop is given by equation 11. The values of β_3 are shown in Fig. 18, for $y_o > 1$.

For an MHD generator, an efficiency ϵ_g may be defined as the ratio of the net power produced to the flow work; that is

$$\epsilon_g = -P_a^+/\Delta p_a Uh \qquad (26)$$

These efficiencies have been calculated from equations 21 and 25 and are shown in Figs. 19–22, for aspect ratios of 1, 2, 3, and 6. It is seen that increasing the aspect ratio increases the efficiency, which is to be expected; also as the exponential magnetic field shading e-folding length is increased, the efficiency is also increased. Also shown as the broken line is a constant magnetic field extension with $c-y_o-10$, which is not as effective, in general, as the exponential shading. This is to be expected from the large eddy currents associated with abrupt change in magnetic field at $y=y_o+c$ from a value of B_1 to zero. Note that at the value of η corresponding to maximum efficiency, the power density is not maximum. This is true also of the ideal generator, because maximum efficiency occurs when $\eta=1$, but then the power is zero.

Pump efficiencies corresponding to $\eta > 1$ are defined as follows:

$$\epsilon_p = -\Delta p_a Uh/P_a^+ \qquad (27)$$

and are shown in Figs. 23 and 24, for $\eta = 1.2$. For constant magnetic field extension, there is a maximum with respect to $c-y_o$; no such maximum occurs with the exponential extension, where the pump efficiencies become asymptotic with large y_o. The ideal efficiency, given by η^{-1} is 0.833. As was the case for the generator, exponential magnetic field extensions are more effective than constant magnitude extensions.

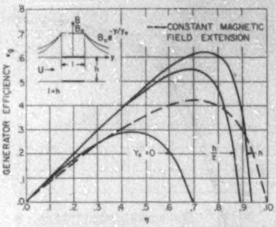


Fig. 19. Actual generator efficiency for channel length equal to channel width

The ohmic heating is the difference between the flow work and the electric power. In terms of the efficiencies this is given by

$$\rho C_v \Delta T = (\epsilon_g - 1)\Delta p_a \qquad (28)$$

for a generator. Thus, small efficiencies lead to large heating. For a pump, this becomes

$$\rho C_v \Delta T = (\epsilon_p^{-1} - 1)\Delta p_a \qquad (29)$$

with a similar conclusion.

Conclusions

In the absence of extensions to the magnetic field, it is evident from the previous sections that the smaller the aspect ratio or the higher the ratio of ϕ_o/UB, the greater the electrical losses. These losses can be corrected by extending the magnetic field both upsteam and downstream of the main magnetic field. It appears that for the integrals of the magnetic field extension ($\int_{y_o}^{\infty} B dY$) which are equal, the constant magnetic field extension is more effective in reducing the electrical losses.

However, the pressure drop or increase is also changed by extending the magnetic field, which in turn affects the efficiency. For a generator it is found that there is a maximum efficiency with respect to η which increases with both increasing aspect ratio and magnetic field extension. It is found that the exponential magnetic field extension is more effective than the constant magnetic field extension for increasing the generator efficiency, since the constant magnitude extension has a greater pressure drop than the exponential extension. This is caused by the larger induced electromotive force by B_o while electromotive force induced by the exponential extension decreases gradually.

For pumping, the efficiency increases with increasing aspect ratio. However, for the constant magnitude extension, the efficiency does not increase monotonically with increasing extension, but has a maximum. The constant magnitude extension that is required to achieve this maximum increases with increasing aspect ratio. The pumping efficiency associated with the exponential magnetic field extension increases monotonically with increasing e-folding length and does not exhibit such a maximum. The exponential extension is also more effective than the constant magnitude extension in increasing the pumping efficiency.

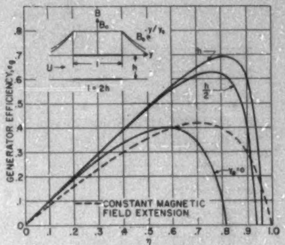


Fig. 20. Actual generator efficiency for channel length equal to two times channel width

If an insulating vane of zero thickness is placed in the center of the channel both upstream and downstream of the main magnetic field, this has the effect of doubling the apparent aspect ratio. If two vanes are placed in the flow, this will treble the apparent aspect ratio, etc. Thus, vanes are an alternate method of increasing the efficiencies.

Appendix I. Electrical Characteristics

Magnetic Field Ends of Electrodes

The boundary conditions in terms of the nondimensional variables are as follows (see Figs. 2, 3) for one quadrant of the channel:

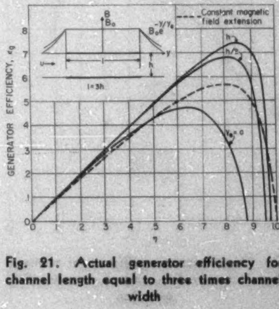


Fig. 21. Actual generator efficiency for channel length equal to three times channel width

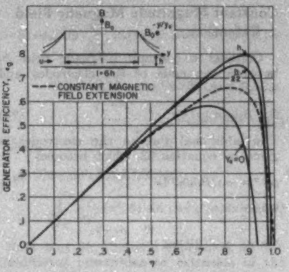


Fig. 22. Actual generator efficiency for channel length equal to six times channel width

$x=0,\ -\infty<y<\infty:\varphi=0$ (from symmetry)

$y=0:\partial\phi/\partial y=0$ (from symmetry)

$x=\pm\pi/2,\ |y|<y_o:\phi=\phi_o$ (specified)

$x=\pm\pi/2,\ |y|>y_o:\partial\phi/\partial x=0\ (j_x=0)$ (30)

To solve equation 3 with boundary conditions given by equation 30, the shaded region of Fig. 2 is mapped into the upper half of the complex w-plane by the transform:

$$w=(1/2)(\sin z+1);\ z=x+iy \tag{31}$$

The upper half of the w-plane is next mapped into a rectangle in the complex z'-plane by means of the following Schwarz-Christoffel transformation which is the incomplete elliptic integral of the first kind (1); see Fig. 3.

$$\begin{aligned} z' &= \frac{1}{2}a[K(1,k)]^{-1}\int_o^w \times \\ &\qquad [w(1-w)(1-k^2w)]^{-1/2}dw \\ &= a[K(1,k)]^{-1}\,sn^{-1}(w^{1/2},k) \end{aligned} \tag{32}$$

where

$$b/a=K'/K \tag{33}$$

$$K'=K(1,k')=K(1,\sqrt{1-k^2}) \tag{34}$$

and

$$k^2=2(1+\cosh y_o)^{-1} \tag{35}$$

The boundary conditions in the z' plane are

$x'=0:\ \phi=0$ (from symmetry)

$x'=a:\ \phi=\phi_o$ (specified)

$y'=0:\ \partial\phi/\partial y'=0$ (from symmetry)

$y'=b:\ \partial\phi/\partial y'=0$ (zero current flow) (36)

The last boundary condition arises from the requirement that the current to the insulator be zero in the region beyond the electrode. Since there is no magnetic field there, the normal gradient of the electric potential must be zero. For the boundary conditions given by the equations in 36, the potential is simply

$$\phi=(\phi_o/a)x' \tag{37}$$

It is convenient to calculate the flux function, ψ, where

$$\begin{aligned} \partial\psi/\partial y' &= \partial\phi/\partial x' \\ \partial\psi/\partial x' &= -\partial\phi/\partial y' \end{aligned} \tag{38}$$

since the current to half the electrode is proportional to the difference in ψ between the ends of the electrodes. Thus,

$$\begin{aligned} \frac{1}{2}J &= \int_o^{Y_o} j\,dY = \sigma\int_o^{Y_o}\left(-\frac{\partial\phi}{\partial x}+UB_o\right)dY \\ &= \sigma UB_oY_o - \int_o^{Y_o}\sigma\frac{\partial\phi}{\partial x'}\,dy' \\ &= \sigma UB_oY_o - \sigma[\psi(a,b)-\psi(a,o)] \end{aligned} \tag{39}$$

From equations 37 and 38, the flux function is

$$\psi=\phi_o(y'/a)$$

The total flux to the electrode is therefore $\psi(a,b)$ which from equation 33 is given by equation 40.

$$\psi(a,b)=\phi_o(b/a)=(K'/K)\phi_o \tag{40}$$

The ideal total current to half the electrode is

$$\begin{aligned} \frac{1}{2}J_i &= \sigma UB_oY_o - \sigma\phi_oY_o/\frac{1}{2}h \\ &= \sigma UB_oY_o - \sigma\varphi_o\pi^{-1}y_o \end{aligned} \tag{41}$$

Thus the difference between the ideal flux and the actual flux is the flux loss

$$\psi_L=\phi_o[(y_o/\pi)-(K'/K)] \tag{42}$$

For values of y_o larger than 0.7, the following asymptotic relation for K'/K may be used with less than 5% error:

$$K'/K=\frac{2\ln 2+y_o}{\pi} \tag{43}$$

while for very small values of y_o

$$K'/K=\frac{\pi/2}{\ln 8-\ln y_o} \tag{44}$$

The total current to one-half of the positive electrode is then

$$\frac{1}{2}J=\sigma[UB_oY_o-(K'/K)\phi_o] \tag{45}$$

The current $(1/2)J$ when multiplied by ϕ_o gives the power to that half electrode. The total actual power is four times that, or

$$P_a=\sigma\eta U^2B_o^2h^2[a-\eta(K'/K)] \tag{46}$$

Equation 46 is plotted as the solid line in Fig. 4.

The maximum power density is obtained by differentiating equation 43 which yields the corresponding value of η as follows:

$$\eta_{a(\max)}=y_oK/(2\pi K') \tag{47}$$

For the ideal generator the maximum power density is obtained when $\eta_{i(\max)}=0.5$, from equation 10. The ratio of the actual maximum power density, evaluated at $\eta_{a(\max)}$, to the ideal maximum power density, is then given by the following equation:

$$P_{a(\max)}/P_{i(\max)}=y_oK/\pi K' \tag{48}$$

The power density and corresponding $\eta_{a(\max)}$ is shown in Fig. 6 for various length channels.

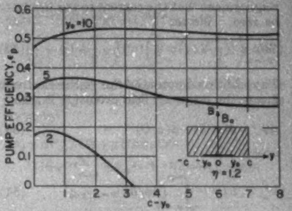


Fig. 23. Pump efficiency for $\eta=1.2$ for constant magnetic field extension

Another interesting quantity is the power "lost," defined as the difference between the ideal power and the actual power. The ratio of the power lost to ideal power is given by the following expression:

$$\begin{aligned} P_L/P_i &= 1-(P_a/P_i) \\ &= a^{-1}\eta(1-\eta)^{-1}[(K'/K)-a] \end{aligned} \tag{49}$$

For very long electrodes, the approximation given by equation 43 may be substituted into equation 49 which yields the following expression:

$$P_L/P_i=2\eta(1-\eta)^{-1}y_o^{-1}\ln 2 \tag{50}$$

Equations 49 and 50 are plotted in Fig. 7, where it is seen that equation 50 is accurate to within 5% for $y_o>0.7$.

Magnetic Fields Extended Beyond Electrodes

When the magnetic field is extended beyond the electrode, because the governing application and boundary conditions are linear, it is necessary only to add to the previous solution another potential ϕ_+ with its associated flux function ψ^+. The boundary conditions for ϕ_+ are as follows:

$x=0,\ -\infty<y<\infty:\phi^+=0$ (51A)

$x=\pm\pi/2,\ |y|<y_o:\phi^+=0$ (51B)

$x=\pm\pi/2,\ |y|>|y_o|:\partial\phi^+/\partial y$
$=UB(y)h/2\pi=\partial\psi/\partial y$ (51C)

$y=0,\ -\frac{\pi}{2}<x<\frac{\pi}{2}:\partial\phi^+/\partial y=0$ (51D)

Boundary condition 51C arises from the

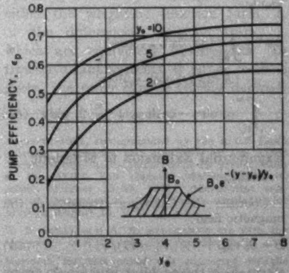


Fig. 24. Pump efficiency for $\eta=1.2$ for exponential magnetic field extension

requirement that the current to the insulator be zero. We set

$$\lim_{y\to\infty} \psi^+(\pi/2, y)=0 \qquad (52)$$

The additional potential flux to the electrode due to the magnetic field extension is then

$$\Delta\psi^+=\psi(a, b)-\psi(a, 0) \qquad (53)$$

where

$$\psi^+(x', b)=-\int_y^\infty (\partial\phi/\partial x)_{x=\pi/2\,dy} \qquad (54)$$

A solution to Laplace's equation which fits the boundary conditions given by the equations in 51 is as follows:

$$\phi^+(x', y')=\sum_{n=1}^{\infty} a_n \sin(n\pi x'/a)\times \cosh(n\pi y'/a) \qquad (55)$$

whence

$$\psi^+(x', y')=\sum_{n=0}^{\infty} a_n \cos(n\pi x'/a)\times \sinh(n\pi y'/a) \qquad (56)$$

Therefore

$$\psi^+(a, 0)=a_o \qquad (57)$$

The value of a_o can be obtained from the value of ψ^+ along (x', b) given by equation 54:

$$a_o=a^{-1}\int_o^a \psi^+(x', b)dx' \qquad (58)$$

Combining equations 53, 54, and 58, the potential flux to the electrode is

$$\Delta\psi^+=\int_{y_o}^{\infty} (\partial\phi/\partial x)_{x=\pi/2}dy- a^{-1}\int_o^a \psi^+(x', b)dx' \qquad (59)$$

Along the line $y=b$,

$$dx'=(a/4K)[w(1-w)(1-l^2w)]^{-1/2}\times \sinh y dy \qquad (60)$$

but

$$w=\frac{1}{2}(1+\cosh y) \qquad (61)$$

Combination of equations 59, 60, and 61 yields

$$\Delta\psi^+=\int_{y_o}^{\infty} (\partial\phi/\partial x)_{x=\pi/2}dy- (K/2)(1+\cosh y_o)^{1/2}\times \int_{y_o}^{\infty} (\cosh y-\cosh y_o)\psi^+(\pi/2, y)dy \qquad (62)$$

Exponential Extension to Magnetic Field

Consider the following extension to the magnetic field:

$$B=B_o \exp\{-[(|Y|-Y_o)/Y_e]\} \qquad (63)$$

where Y_e is the e-folding length. In terms of the transformed co-ordinates, equation 63 becomes

$$B=B_o \exp\left[-\frac{1}{2}\gamma(|y|-y_o)\right] \qquad (64)$$

where $\gamma=2/y_e$ and $y_e=2\pi Y_e/h$. When equation 64 is substituted into equation 51C and integrated, the value of ψ along the insulator is obtained. This is then substituted into equation 62 to obtain the additional flux, as follows:

$$\Delta\psi_+ = -(2\phi_o\beta_1)/(\eta\pi) \qquad (65)$$

where β_1 is defined as

$$\beta_1=\gamma^{-1}\left\{1-(2^{1/2}kK)^{-1}\times \int_{y_o}^{\infty} (\cosh y-\cosh y_o)^{-1/2}\times \exp\left[-\frac{1}{2}\gamma(y-y_o)\right]dy\right\} \qquad (66)$$

Values of β_1 are shown in Fig. 8 or various values of y_o and β_1, where it is evident that β_1 is insensitive to the channel length as the channel length increases beyond a value of y_o equal to 2.

For long electrodes ($y_o>1$), the integral may be simplified by letting

$$\exp\left[-\frac{1}{2}(y-y_o)\right]=\sin x' \qquad (67)$$

so that the second term in the brackets of equation 66 becomes

$$-2\pi^{-1}\int_o^{\pi/2} (\sin x')^\gamma dx' =\pi^{-1/2}\Gamma\left(\frac{1}{2}\gamma+\frac{1}{2}\right)/\Gamma\left(\frac{1}{2}\gamma+1\right) \qquad (68)$$

and equation 66 becomes

$$\beta_1=\gamma^{-1}\left[1-\pi^{-1/2}\Gamma\left(\frac{1}{2}\gamma+\frac{1}{2}\right)/\Gamma\left(\frac{1}{2}\gamma+1\right)\right] \qquad (69)$$

Finally, for a magnetic field which extends to $y=\infty$, the results of equation 68 may be used by letting $\gamma\to 0$. Then

$$\lim_{\gamma\to 0}\Gamma\left(\frac{1}{2}\gamma+\frac{1}{2}\right)/\Gamma\left(\frac{1}{2}\gamma+1\right) =\pi^{1/2}(1-\gamma\ln 2) \qquad (70)$$

The additional power due to the extension of the magnetic field is

$$\Delta P^+=2\sigma\eta\pi^{-1/2}U^2B_o^2h^2\beta_1 \qquad (71)$$

From equations 70 and 49 the ratio of additional power due to the extension of the magnetic field, to that lost by termination at the electrode ends, is given by

$$\Delta P^+/P_L=2\pi^{-1}\eta^{-1}\beta_1[(K'/K)-a]^{-1} \qquad (72)$$

For a long channel and long magnetic field extension, use of equation 70 in equation 66 and equation 72 yields

$$\Delta P^+/P_L=\eta^{-1} \qquad (73)$$

These results show that for $o<\eta<1$, the additional power obtained by a very large extension of the magnetic field is larger than the loss associated with the termination of the constant magnetic field at the end of the electrode. For $\eta>1$, the additional power input tends to reduce the net power required for a given pressure increase.

Constant Magnitude Magnetic Field Extension

Consider a constant magnetic field B_1 between the ends of the electrodes and $Y=C$:

$$B=B_1; y_o<|y|<c \qquad (74)$$

The additional flux to the electrodes is given by equation 62, which becomes

$$\Delta\psi^+=-UB_1h\beta_2/2\pi \qquad (75)$$

where

$$\beta_2=c-y_o-(2^{1/2}kK)^{-1}\int_{y_o}^{c-y_o}\times (\cosh y-\cosh y_o)^{-1/2}(c-y)dy \qquad (76)$$

For values of $y_o>1$, this becomes

$$\beta_2=4\pi^{-1}\left\{-\frac{1}{2}x_o'(c-y_o)+\int_{x_o'}^{\pi/2} \ln \sin x'dx'\right. \qquad (77)$$

where

$$x_o'=\sin^{-1}\exp\left[-\frac{1}{2}(c-y_o)\right] \qquad (78)$$

The function β_2 is shown in Fig. 13. As $c\to\infty$ the current is the same as that given by equation 65.

Appendix II. Pressure Change

The average pressure change is given by the total Lorentz force, averaged over the cross section:

$$\Delta p_a=h^{-1}\int_{-\infty}^{\infty}\int_{-h/2}^{h/2} j_x B dX dY \qquad (79)$$

Upon making use of the generalized Ohm's law, and integration with respect to X, this becomes

$$\Delta p_a=4\sigma h^{-1}\int_{Y_o}^{\infty}\phi\left(\frac{h}{2}, Y\right)B(Y)dY- 2\sigma\int_{Y_o}^{\infty} UB^2(Y)dY+\Delta p_i \qquad (80)$$

where the ideal pressure change is given by

$$\Delta p_i=2(\eta-1)UB_o^2Y_o \qquad (81)$$

Equation 80 shows the interesting result that the pressure change is equal to the ideal pressure change in the absence of any extension of the magnetic field even though there is a power loss. To evaluate the first term of equation 80 note that for $Y>Y_o$, the current to the wall is zero, hence $B=(\partial\phi/\partial X)/U$. The first term then becomes

$$(4\sigma/Uh)\int_{Y_o}^{\infty}\phi\left(\frac{h}{2}, Y\right)\frac{\partial\phi}{\partial X}\left(\frac{h}{2}, Y\right)dY \qquad (82)$$

or

$$-(4\sigma/Uh)\int_a^o \phi(b, x')\frac{\partial\phi}{\partial y'}(b, x')dx' \qquad (83)$$

The values of ϕ and its normal derivative are obtained from equations 37 and 55; equation 83 then becomes

$$-(4\sigma\phi_0/Uh)\sum_{n=1}^{\infty} a_n(-1)^n \sinh(n\pi K'/K)+$$

$$(2\sigma\pi/Uh)\sum_{n=1}^{\infty} na_n^2 \cosh(n\pi K'/K)\times$$

$$\sinh(n\pi K'/K) \quad (84)$$

where

$$a_n=2[n\pi \sinh(n\pi K'/K)\times$$

$$\int_0^a \frac{\partial\phi}{\partial y'}(x', b)\sin(n\pi x'/a)dx' \quad (85)$$

Note that an alternate expression for the additional power due to extension of the magnetic field is given by

$$\Delta P^+=-4\sigma\phi_0\int_0^{Y_0}\left(\frac{\partial\phi}{\partial x}\right)dY$$

$$=-4\sigma\phi_0\int_0^b \frac{\partial\phi}{\partial x'}(a, y')dy' \quad (86)$$

which becomes

$$\Delta P^+=-4\sigma\phi_0\sum_{n=1}^{\infty} a_n(-1)^n \sinh(n\pi K'/K) \quad (87)$$

Thus the first term in equation 84 is just $\Delta P^+/Uh$. For very long electrodes ($y_0>1$), the actual pressure change becomes

$$\Delta p_a=\Delta P^+/Uh+16\pi^{-3}\sigma UhB_0^2\times$$

$$\sum_{n=2,4,6\ldots}^{\infty} nA_n^2-2\sigma U\int_{Y_0}^{\infty} B^2dY+\Delta p_i \quad (88)$$

where

$$A_n=n^{-1}\int_0^{\pi/2}(B_0 \sin x')^{-1}B(x')\times$$

$$\cos(x')\sin(nx')dx' \quad (89)$$

and x' is defined by

$$x'=\sin^{-1}\exp\left[-\frac{1}{2}(y-y_0)\right] \quad (90)$$

For exponential extension of the magnetic field, equation 88 becomes

$$\Delta p_a=\Delta P^+/Uh+\sigma UhB_0^2\times$$

$$\left[16\pi^{-3}\sum_{n=2,4,6}^{\infty} nA_n^2-(\gamma\pi)^{-1}\right]+\Delta p_i \quad (91)$$

where the coefficient A_n is given by the integration of equation 89 as follows:

$$A_n=2^{-(\gamma+1)}\gamma^{-1}\pi(-1)^{n/2}\Gamma(\gamma+1)$$

$$\Gamma^{-1}\left(\frac{1}{2}\gamma+1+\frac{1}{2}n\right)\Gamma^{-1}\left(\frac{1}{2}\gamma+1-\frac{1}{2}n\right) \quad (92)$$

For extension of the constant magnetic field of B_0 to c, the pressure change is given by

$$\Delta p_a=\Delta P^+/Uh+\sigma UhB_0^2\times$$

$$[16\pi^{-3}\sum nA_n^2-\pi^{-1}(c-y_0)]+\Delta p_i \quad (93)$$

The functions in the brackets of equations 91 and 92, which for convenience will be designated β_2, are shown in Fig. 18. For large extensions of the magnetic field, Δp_a becomes asymptotic as does P_a^+: hence there is a limiting value to the pressure change.

References

1. METHODS OF THEORETICAL PHYSICS (book), Philip M. Morse, Herman Feshbach. McGraw-Hill Book Company, Inc., New York, N. Y., 1953, p. 1251.

2. END EFFECTS IN MAGNETOHYDRODYNAMIC CHANNEL FLOW, F. Fishman. *Research Report 78, Contract AF 04(647)-275*, AVCO Everett Research Laboratory, Everett, Mass., June 1959.

Signal Flow Graph Theory for Linear Time-Variable Systems

E. B. STEAR
NONMEMBER AIEE

A. R. STUBBERUD
NONMEMBER AEE

Summary: Techniques here presented for the analysis of linear, time-variable systems through signal flow graph concepts are based on the combinatorial properties of linear differential equations. The concepts of stability and sensitivity of linear, time-variable systems, which are described by linear differential equations, are discussed. A linear, time-variable feedback system is analyzed to aid in clarification.

CONCEPTS AND RESULTS on signal flow graph theory[1] are generalized in such a way that the signal flow graph will be equally useful for analyzing time-variable and time-invariant systems. Since the concept of the Laplace transform is not highly useful and elements are not commutable in time-variable systems, the reduction of a time-variable signal flow graph becomes complex and, in general, is most easily carried out in the time domain, directly in terms of the differential equations of the individual transmittances. The classical reduction method in time-variable cases involves working with weighting functions.[2-5] However, this scheme has a serious drawback in that, whenever it is necessary to find an inverse transmittance (stability of a self-loop), a Volterra integral equation must be solved. This difficult problem can be avoided by working directly with differential equations,[4] as will be seen.

The rules for drawing a signal flow graph for a general linear system are generalizations of the same rules presented in reference 1, and can be stated:

1. Signals travel along branches only in the directions of the arrows.
2. A signal traveling along any branch is convolved with the weighting function of that branch.
3. The value of the variable represented by any node is the sum of all signals entering the node.
4. The value of the variable represented by any node is transmitted on all branches leaving that node.

The transmittances which represent the various branches of the flow graphs will, in this paper, be assumed stable in some sense. Discussions of stability can be found in references.[8-10] In addition, the transmittances shall be limited to those which can be described by weighting functions of the form

$$W(t, \tau)=\sum_{i=1}^{n} a_i(\tau)q_i(t)+b_n(t)\delta(t-\tau)t \geq \tau$$

$$=0;\ t<\tau \quad (1)$$

where $W(t, \tau)$ is the particular solution $y(t)$ of the equation

$$\sum_{l=0}^{n} a_l(t)\frac{d^l y}{dt^l}=\sum_{j=0}^{n} b_j(t)\frac{d^j x}{dt^j}\quad a_n(t)=1 \quad (2)$$

if $x=\delta(t-\tau)$.

In practice, the transmittances of a linear system may be represented in either the weighting function form, 1, or the differential equation form, 2, which are equivalent descriptions. In time-invariant systems, if either form is given, the other can be readily derived;.

Paper 61-1051, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted December 21, 1960; made available for printing August 15, 1961.

E. B. STEAR and A. R. STUBBERUD are with the University of California, Los Angeles, Calif.

Research for this paper was supported by the United States Air Force under Contract No. *AF 49(638)-438*, monitored by the Air Force Office of Scientific Research of the Air Research and Development Command.

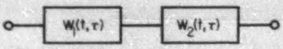


Fig. 1. Cascade transmittances

however, this is not the case in time-variable systems. As will be shown later, a form 1 can be transformed into a form 2; the converse, however, is not true.

In order to apply the concepts of signal flow graph theory to linear, time-variable systems, certain combinatorial techniques—analogous to addition, subtraction (the additive inverse), multiplication, and division (the multiplicative inverse)—must be available. If weighting functions could be easily determined for linear, time-variable differential equations and if inverse weighting functions (references 4, 5,11) could be easily determined, the convolution integral would be an adequate tool for reducing any linear signal flow graph. However, as this is not the case, less elegant techniques must be used.

The next four sections of this report treat the problems of: (1) development of differential equations from given weighting functions; (2) cascade combination of transmittances; (3) parallel combination of transmittances; and (4) determination of inverse transmittances. Using the concepts which are developed, a linear signal flow graph may be reduced to an essential diagram containing transmittances described by single differential equations, or, if desired, the system can be reduced to a single differential equation.

Differential Equations from Weighting Functions[3,11]

The weighting function of a linear system which is described by a differential equation of the form in equation 2 can be defined as the response of the system when the driving function x is a unit impulse at time $t=\tau$. This response will have the form given in equation 1. In practice, it is difficult to determine the weighting function of a linear, time-variable differential equation because the independent solutions, in general, cannot be expressed in a closed form nor in terms of known tabulated transcendental functions. On the other hand, if a time-variable weighting function is given, the differential equation which it satisfies can be constructed quite easily.

The need for construction of a differential equation from a given weighting function arises in two cases in the reduction of a time-variable signal flow graph. First, if some of the graph transmittances are given in form 1 and the rest in form 2, before cascade or parallel combinations of any two of these transmittances can be performed, they will have to be in the same form. Since determination of the weighting function of a differential equation is very difficult, the obvious step is to change the weighting functions to differential equations. Second, as will be shown, when the inverse of a transmittance is required, it can be obtained very simply in differential equation form.

The q's of a weighting function in the form of equation 1 satisfy equations of

$$\sum_{l=0}^{n} a_l(t)\frac{d^l q_i}{dt^l}=0$$

$$i=1,2\ldots n \qquad (3)$$

Equations 3 represent a system of n equations in the $n+1$ unknown a's. One of the unknown a's can be chosen with no loss of generality. By arbitrarily choosing $a_n(t)=1$, equation 3 can be solved for the remaining unknown a's. Since the determinant of the system equations is the Wronskian of the solutions q of the differential equation, solutions to equations 3 can always be obtained. Once the a's are determined, the b's can also be obtained.

The first step in determining the b's of the differential equation is to determine a set of functions which shall be denoted $F_i(t)$.[3,11] These functions are defined by

$$F_0(t)=b_n(t)$$

$$F_i(t)=\frac{\partial^{i-1}}{\partial_t{}^{i-1}}[W(t,\tau)]\Big|_{t=\tau^+}$$

$$i=1,2\ldots n \qquad (4)$$

where

$$\frac{\partial^j}{\partial t^j}[W(t,\tau)]\Big|_{t=\tau^+}$$

is the amplitude of the discontinuity in the jth derivative of the weighting function at $t=\tau$.

The b's of equation 2 can now be determined from the relationship[6]

$$b_{n-i}(t)=\sum_{k=0}^{i}\sum_{s=0}^{i-k}\binom{n+s-i}{n-i}a_{n-i+k+s}(t)\times$$

$$\frac{d^sF_k}{dt^s} \quad i=1,2,2\ldots n \qquad (5)$$

where

$$\binom{n+s-i}{n-i}$$

is the binomial coefficient.

The method of determining a differential equation from its known weighting function is then complete; however, a word is in order concerning differentiability conditions on the β's and q's of the given weighting function. Examination of equations 4 and 5 reveals that the q's must have $2n$ continuous derivatives and the β's must have n continuous derivatives. These conditions guarantee the continuity of the a's and b's.

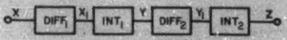


Fig. 2. Cascade transmittances, with differential and integral parts represented

Cascade Combination of Nonstationary Transmittances[12]

As was pointed out earlier, a transmittance may be equally well described by either a weighting function or a differential equation. Suppose in the reduction of a signal flow diagram, two cascade transmittances are to be replaced by a single equivalent transmittance. Two possible situations may arise: (1) both transmittances may be described by weighting functions, or (2) one or both may be described by a differential equation. The first situation can be quite easily handled by convolving the two transmittances. Consider Fig. 1 in which

$$W_1(t,\tau)=\sum_{i=1}^{n}\eta_i(\tau)u_i(t)+b_{n_1}(t)\delta(t-\tau)$$

$$W_2(t,\tau)=\sum_{j=1}^{m}\gamma_j(\tau)v_j(t)+b_{n_2}(t)\delta(t-\tau) \qquad (6)$$

The weighting function of a single equivalent transmittance is given by the convolution integral

$$W(t,\tau)=\int_{\tau}^{t}W(t,\theta)W_1(\theta,\tau)d\theta \qquad (7)$$

Substituting 6 into 7, carrying out the indicated integration, and then forming products of functions in t and τ produces the over-all weighting function

$$W(t,\tau)=\sum_{j=1}^{m}v_j(t)\Bigg[\gamma_j(\tau)b_{n_1}(\tau)-$$

$$\sum_{i=1}^{n}\eta_i(\tau)F_{ij}(\tau)\Bigg]+\sum_{i=1}^{n}\times$$

$$\Bigg[u_i(t)b_{n_2}(t)+\sum_{j=1}^{m}F_{ij}(t)v_j(t)\Bigg]\eta_i(\tau)+$$

$$b_{n_1}(t)b_{n_2}(t)\delta(t-\tau) \qquad (8)$$

where $F_{ij}(\alpha)$ is the indefinite integral

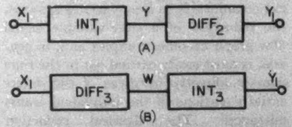


Fig. 3. Equivalent time-variable transmittances

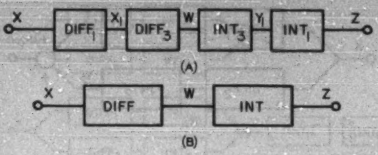


Fig. 4. Transmittances equivalent to Fig. 2

$$\int^{\alpha} u_i(\theta)\gamma_j(\theta)d\theta$$

Equation 8 obviously has the same form as equation 1 and has a maximum order $(m+n)$.

Suppose now that the second situation should occur, i.e., one or both transmittances are in differential equation form. In this case, the first step is to get both transmittances into the differential equation representation by the method in the preceding section. Fig. 2 represents a block diagram of the cascade transmittances in which "Diff" and "Int" represent the differential and integral parts of each transmittance, respectively. Obviously, any differential equation can be represented in this manner.

The relationships between variables in Fig. 2 are given by

$$\sum_{j=0}^{n} b_j x^{(j)} = x_1 = \sum_{i=0}^{n} a_i y^{(i)}$$

$$\sum_{l=0}^{m} d^l y^{(l)} = y_1 = \sum_{k=0}^{m} c_k z^{(k)} \tag{9}$$

where

$$y^{(i)} = \frac{d^i y}{dt^i}$$

$$x^{(j)} = \frac{d^j x}{dt^j}$$

and so forth.

Suppose now the two inner blocks of Fig. 2 are operated upon as shown in Fig. 3, i.e., the integral and differential operations are interchanged. Because of the noncommutability of time-variable elements

$\text{Diff}_2 \neq \text{Diff}_1$

$\text{Int}_1 \neq \text{Int}_2$

Figs. 3(A) and 3(B) are equivalent between input and output but not at the intermediate point. Equations relating the variables in Fig. 3(B) are

$$\sum_{\alpha=0}^{m} g_\alpha x_1^{(\alpha)} = W$$

$$\sum_{\beta=0}^{n} h_\beta y_1^{(\beta)} = W \tag{10}$$

The problem here is to find the g_α and h_β in terms of the known a_i and d_l. Substituting from equation 9 for x_1 and y_1, both in terms of y, produces the relationship

$$\sum_{\alpha=0}^{m} \sum_{i=0}^{n} \sum_{c=0}^{\alpha} g_\alpha \binom{\alpha}{c} a_i^{(\alpha-c)} y^{(i+c)}$$

$$= \sum_{\beta=0}^{n} \sum_{l=0}^{m} \sum_{\delta=0} h_\beta \binom{\beta}{\delta} d_l^{(\beta-\delta)} y^{(l+\delta)} \tag{11}$$

By equating like derivatives of y in equation 11, a set of $m+n+1$ simultaneous algebraic equations in the $m+n+2$ unknown, g_α and h_β are obtained. Arbitrarily choosing $h_n=1$ (without loss of generality) allows these equations to be solved for the remaining $m+n+1$ unknowns. Fig. 2 then can be replaced by its equivalent of Fig. 4(A). Fig. 4(A) can be further reduced by combining the differential and integral portions of the system. The result is a single $(n+m)$th-order linear differential equation relating x and z, the relation which was desired. This equation has the form

$$\sum_{\beta=0}^{n} \sum_{k=0}^{m} \sum_{f=0}^{\beta} \binom{\beta}{f} h_\beta c_k^{(\beta-f)} z^{(k+f)} =$$

$$\sum_{\alpha=0}^{m} \sum_{j=0}^{n} \sum_{r=0}^{\alpha} \binom{\alpha}{r} g_\alpha b_j^{(\alpha-r)} x^{(j+r)} \tag{12}$$

Equations 11 and 12 or the convolution integral will serve to combine cascade transmittances into a single equivalent transmittance.

Inverse Transmittances

The third technique which is necessary for reducing a nonstationary signal flow diagram is that of determining inverse transmittances. This problem is represented diagrammatically in Fig. 5(A), i.e., the inverse of a transmittance is such that a cascade of a transmittance and its inverse leaves the input signal x unaltered. In this entire section, all initial conditions are assumed to be zero. In terms of a weighting function $W(t, \tau)$ and its inverse $[W(t, \tau)]^{-1}$, the problem amounts to solving the integral equation

$$\delta(t-\tau) = \int_{\tau}^{t} W(t, \theta)[W(\theta, \tau)]^{-1} d\theta \tag{13}$$

for the unknown inverse $[W(t, \tau)]^{-1}$. Solution of this equation is, in general, difficult. Therefore, the weighting function representation of transmittances will be discarded in favor of the differential equation representation.

To find the differential equation of the inverse system, consider the cascade transmittances of Fig. 5(B), in which the transmittances of Fig. 5(A) have been separated into integral and differential operators. The operator Diff₁ has the form

$$\sum_{j=0}^{n} b_j x^{(j)} = z \tag{14}$$

and Int₁ has the form

$$z = \sum_{i=0}^{n} a_i y^{(i)} \tag{15}$$

If the weighting function corresponding to Int₁ is called $W_1(t, \tau)$, then

$$y(t) = \int_{-\infty}^{t} W_1(t, \tau) z(\tau) d\tau \tag{16}$$

It can be shown[13] that operating on $y(t)$ with a differential operator of the following form:

$$\sum_{i=0}^{n} a_i \frac{d^i}{dt^i} \tag{17}$$

produces $z(t)$; thus, equation 17 is the inverse of Int₁.

If Diff₂ is chosen with this differential form, Fig. 5(B) can be reduced to Fig. 6(A). Suppose that the weighting function corresponding to Int₂ is represented

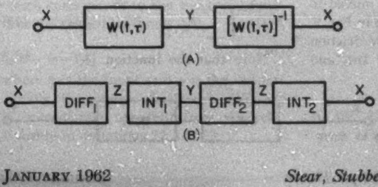


Fig. 5 (left). Cascade of a transmittance and its Inverse

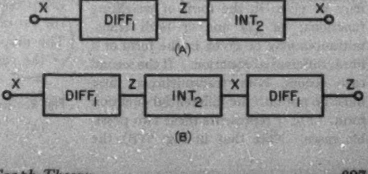


Fig. 6 (right). Reduction of Fig. 5(B)

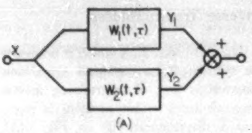
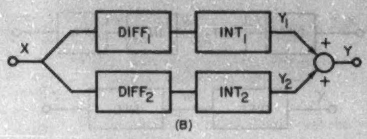


Fig. 7 (left and right). Parallel transmittances

A—Weighting function representation

B—Differential equation representation



by $W_2(t, \tau)$. Since $W_2(t, \tau)$ is to be the inverse of Diff$_1$, the convolution integral

$$x(t)=\int_{-\infty}^{t} W_2(t,\tau)z(\tau)d\tau \tag{18}$$

must hold.

Suppose now the differential operator Diff$_1$ is applied to this integral as indicated in Fig. 6(B). Then from equation 14

$$\sum_{j=0}^{n} b_j x^{(j)}=z(t)=\sum_{j=0}^{n} b_j \frac{\partial^j}{\partial t^j}\times \left[\int_{-\infty}^{t} W_2(t,\tau)z(\tau)d\tau\right] \tag{19}$$

$W_2(t, \tau)$ satisfies this equation, if it satisfies

$$\sum_{j=0}^{n} b_j \frac{\partial^j}{\partial t^j}[W_2(t,\tau)]=\delta(t-\tau) \tag{20}$$

Int$_2$ can be represented, therefore, by the equation

$$z=\sum_{j=0}^{n} b_j x^{(j)} \tag{21}$$

Now the final result can be stated. If a transmittance is represented by the equation

$$\sum_{j=0}^{n} b_j x^{(j)}=\sum_{i=0}^{n} a_i y^{(i)} \tag{22}$$

where x is the input and y is the output, then its inverse can be represented by the equation

$$\sum_{i=0}^{n} a_i y^{(i)}=\sum_{j=0}^{n} b_j x^{(j)} \tag{23}$$

where y is the input and x is the output.

Parallel Combination of Transmittances

There are two possible cases which might arise in the combination of parallel transmittances: (1) both transmittances may be given in the form of weighting functions; or (2) one or both transmittances may be given in the form of a linear differential equation. If the second case occurs, both transmittances first must be put into the differential equation form. Fig. 7 represents these two possible cases. Note that in Fig. 7(B), the differential equations have been divided into their differential and integral parts.

Consider case 1, represented by Fig. 7(A) in which

$$W_1(t,\tau)=\sum_{i=1}^{n} \beta_i(\tau)q_i(t)+b_{n_1}(t)\delta(t-\tau)$$

$$W_2(t,\tau)=\sum_{j=1}^{m} \alpha_j(\tau)u_j(t)+b_{n_2}(t)\delta(t-\tau) \tag{24}$$

It is seen from the figure that

$$y=y_1+y_2=\int_{-\infty}^{t}[W_1(t,\tau)+W_2(t,\tau)]x(\tau)d\tau=\int_{-\infty}^{t} W(t,\tau)d\tau \tag{25}$$

where

$$W(t,\tau)=W_1(t,\tau)+W_2(t,\tau) \tag{26}$$

$$W(t,\tau)=\sum_{i=1}^{n} \beta_i(\tau)q_i(t)+\sum_{j=0}^{m} \alpha_j(\tau)u_j(t)+[b_{n_1}(t)+b_{n_2}(t)]\delta(t-\tau)$$

Obviously, redefinition of the terms in equation 26 will produce a weighting function of the form in equation 1 of $(m+n)$th order.

Now consider case 2, in which the two transmittances are given in differential equation form and represented in Fig. 7(B). The problem is to reduce the system in Fig. 7(B) to the form in Fig. 8 by proper application of the techniques developed in earlier sections of this paper.

The reduction is best illustrated by Fig. 9(A) through 9(G). Step 1 indicates that four operators are cascaded with the original system. Note that (Int)$^{-1}$ and $[(\text{Int}_2)']^{-1}$ are differential operators and are the respective inverses of Int$_1$ and (Int$_2$)′. Int$_1$ apparently is the integral portion of the first transmittance. (Int$_2$)′ will be defined subsequently. Note that addition of these operators leaves y unchanged.

Since the system is linear the differential operator (Int$_1$)$^{-1}$ can be moved to the left of the summer as shown in Fig. 9. The properties of inverse allow deletion of the cascade combination of Int$_1$ and (Int$_1$)$^{-1}$ from the upper branch. Note that the output of the summer is no longer y.

In step 3, the integral operator Int$_2$ and the differential operator (Int$_1$)$^{-1}$ are interchanged by the method described in the section on cascade combinations of transmittances. The new differential and integral operators are then designated $[(\text{Int}_1)^{-1}]'$ and (Int$_2$)′, respectively. This makes obvious the definition of (Int$_2$)′ mentioned in connection with step 1.

Step 4 shows that $[(\text{Int}_2)']^{-1}$ has been moved to the left of the summer, thus removing (Int$_2$)′ from the lower branch. Upper and lower branches now contain only differential operators while the summer is followed by integral operators only. The latter can be combined as shown in step 5. Since the sum of two differential operators, operating on x, can be represented by a single operator operating on x, the system can be reduced as in Fig. 9(G), at which point the reduction is complete. Two differential equations can be differenced if the summer in Fig. 9 is replaced by a differencing circuit and if an identical reduction technique is then applied.

Reduction of a Self-Loop

One concept which can now be developed is that of reducing a self-loop to an equivalent open-loop transmittance. Consider the self-loop of Fig. 10(A). The transmittances from the input node to node 2 and from node 2 to node 3 are unity gains and have weighting functions equal to $\delta(t-\tau)$. The weighting function of the loop is denoted by $W_{22}(t, \tau)$. Now solving for x_2

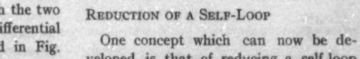
$$x_2(t)=\int_{-\infty}^{t}\delta(t-\tau)x_2(\tau)d\tau=x_1(t)+\int_{-\infty}^{t} W_{22}(t,\tau)x_2(\tau)d\tau \tag{27}$$

which can be rewritten in the following form:

$$x_1(t)=\int_{-\infty}^{t}[\delta(t,\tau)-W_{22}(t,\tau)]x_2(\tau)d\tau \tag{28}$$

Note that the function $[\delta(t-\tau)-W_{22}$-

Fig. 8. Reduced form of Fig. 7(B)



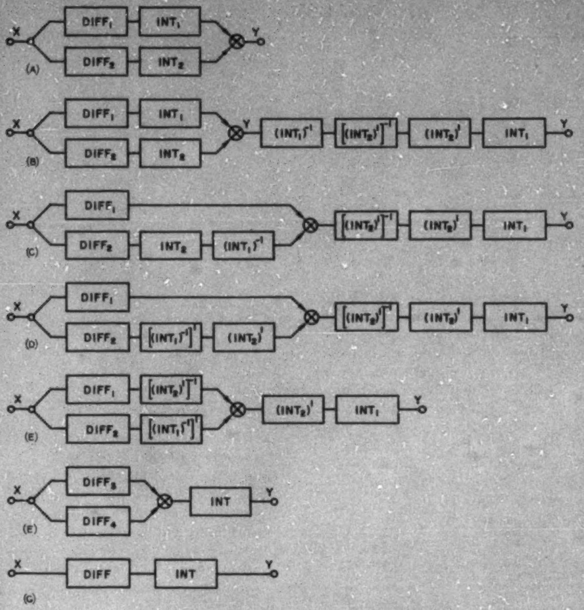


Fig. 9. Reduction of parallel transmittances

A—Original system
B—Step 1
C—Step 2
D—Step 3
E—Step 4
F—Step 5
G—Reduced system

$(t, \tau)]$ can be redefined as a new weighting function $W(t, \tau)$. If $[W(t, \tau)]^{-1}$ is applied to both sides of equation 28, the result is

$$x_2(t) = \int_{-\infty}^{t} [W(t, \tau)]^{-1} x_1(\tau) d\tau \tag{29}$$

Fig. 8(B) is then equivalent to Fig. 8(A). An expression for the transmittance $[W(t, \tau)]^{-1}$ remains to be determined. As $[\delta(t-\tau)-W(t, \tau)]$ is the difference of two transmittances, the parallel combination method of the previous section can be used to determine a differential equation equivalent to $[\delta(t-\tau)-W(t, \tau)]$. It should be kept in mind that the differential equation which corresponds to the weighting function $\delta(t, \tau)$ has the form

$$x = y \tag{30}$$

where x is the input and y is the output. Finally, the inverse of $[\delta(t-\tau)-W(t, \tau)]$ is immediately apparent from the relationship in equations 22 and 23.

Stability[8]

DEFINITION 1

A system is said to be stable if, and only if, every bounded input gives rise to a bounded output.

The concept of the boundedness of a system output, described by a differential equation, can now be explained.

DEFINITION OF BOUNDEDNESS

Given a differential equation

$$y^{(n)} = f(t, y^{(1)} \ldots y^{(n-1)}) \tag{31}$$

A solution $y(t) = y(t;\ t_0, n_0 \ldots n_{n-1})$ is called "bounded at the right" if $|y^{(h)}| \leq M$, M finite, for all $t \geq t_0$ and $h = 0, 1 \ldots n-1$. In defining boundedness, the n's represent the values of the initial conditions at time t_0. It should be noted that equation 2 can be written in the form of 31. A definition analogous to the foregoing definition holds for "bounded at the left."

THEOREM 1

Given the nth-degree algebraic equation

$$\lambda^n + \sum_{i=0}^{n-1} c_i \lambda^i = 0 \tag{32}$$

in which the c_i are real constants and λ_i, $i = 1 \ldots n$ are solutions of the equation; and given the nth-order linear differential equation

$$y^{(n)} + \sum_{i=0}^{n-1} a_i(t) y^{(i)} = F(t) \tag{33}$$

Then if $a_i(t) \rightarrow c_i$ for $i = 0, 1 \ldots n-1$, $F(t) \rightarrow K$ (a finite constant) as $t \rightarrow +\infty$ and if $Re(\lambda_i) < 0$ for $i = 1 \ldots n$, then all solutions of 33 verify the relations $y(t) \rightarrow K/C_0$ and $y^{(r)}(t) \rightarrow 0$ for $r = 1 \ldots n$ as $t \rightarrow +\infty$. If $C_0 \pm 0$, $Re(\lambda_i) \leq 0$ for $i = 1 \ldots n$ and the roots with $Re(\lambda_i) = 0$ are simple and if

$$\int^{+\infty} |c_i - a_i(t)| dt < +\infty \text{ for } i = 0, 1 \ldots n-1$$

and

$$\int^{+\infty} |K - F(t)| dt < +\infty$$

then all solutions of 33 are bounded as $t \rightarrow +\infty$. The same holds if $C_0 = 0$ provided $K = 0$.

Thus equation 2 can be written in the form of 33 if the definition

$$F(t) = \sum_{j=0}^{n} b_j(t) \frac{d^j x}{dt^j} \tag{34}$$

is used. It should be noted that the condition $F(t) \rightarrow K$ requires that $b_0(t) \rightarrow K_1$, $x(t) \rightarrow K_2$ and

$$\frac{d^j x}{dt^j} \rightarrow 0,\ j = 1 \ldots n \text{ as } t \rightarrow +\infty$$

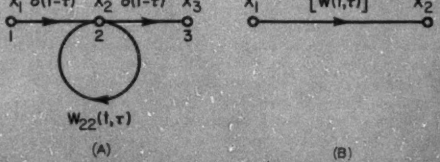


Fig. 10. Signal flow graph of self-loop

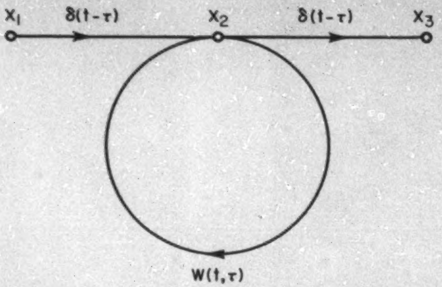


Fig. 11. Feedback system

(K_1 and K_2 are finite constants) and also that the $b_j(t)$ for $j=1 \ldots n$ be bounded as $t \to +\infty$.

Sensitivity

One measure of the quality in a feedback system is its sensitivity to parameter variations. The sensitivity of an over-all gain T with respect to a given parameter K can be defined[1] by the equation

$$E_K{}^T = \frac{dT/T}{dK/K} \qquad (35)$$

In other words, the sensitivity of T with respect to K is obtained by dividing the percentage change in T by that in K, which changes T, but all changes are differentially small.

A factor which complicates using equation 35 to determine the sensitivity of a time-invariant linear system has been pointed out by Truxal.[1] In a general time-invariant system, the system transmittance T is a function of the complex frequency s (assuming the transmittance has been Laplace transformed). The sensitivity $E_K{}^T$ will, therefore, also be a function of s. This means that variation in sensitivity results in varying the nature of the transient response. The significance of this dependence on complex frequency s is, to the authors' knowledge, relatively unknown. One result is that, generally, equation 35 is applied only to frequency independent gains T.

Suppose, however, a frequency dependent gain has been obtained for a linear, stationary system. What can be done with it? First, an inverse transform can be applied and the equivalent time-varying sensitivity obtained which describes the system sensitivity for all time. In many cases, however, the only information desired is the steady-state sensitivity of the gain function.

The concept of a time-varying sensitivity can be extended directly to time-varying systems. Even if a system has no dynamics, its sensitivity will be a function of time in the time-varying case. If the system has dynamics, then it is necessary to solve a time-varying linear differential equation to determine the sensitivity for all time. The steady-state sensitivity can be determined, however, by applying theorem 1 of the section on stability to the differential equation describing the sensitivity. An example will illustrate steady-state sensitivity as well as some other concepts discussed in previous sections.

Example

In the feedback system, Fig. 11, $W(t, \tau)$ is the weighting function corresponding to the second-order linear differential equation

$$y^{(2)} + a_1(t)y^{(1)} + a_0(t)y = b_1(t)x^{(1)} + b_0(t)x \qquad (36)$$

where

$$a_1(t) = 2 + e^{-t}$$

$$a_0(t) = 2 + e^{-2t}$$

$$b_1(t) = e^{-t}$$

and

$$b_0(t) = 1$$

The stability of this system and its steady-state sensitivity to perturbations of $a_0(t)$ should be examined.

According to equations 28 and 29,

$$x_2 = \int_{-\infty}^{t} [\delta(t-\tau) - W(t, \tau)]^{-1} x_1(\tau) d\tau \qquad (37)$$

The first step in the analysis, then, is to form the differential equation equivalent to the weighting function $[\delta(t-\tau) - W(t, \tau)]$. Recalling that the differential equation equivalent to the weighting function $\delta(t-\tau)$ is merely $x=y$ and applying the combinatorial techniques indicated in Fig. 9, the differential equation corresponding to $[\delta(t-\tau) - W(t, \tau)]$ is

$$x_1{}^{(2)} + a_1(t)x_1{}^{(1)} + a_0(t)x_1 = x_3{}^{(2)} + [a_1(t) - b_1(t)]x_2{}^{(1)} + [a_0(t) - b_0(t)]x_3 \qquad (38)$$

where x_1 and x_2 are as indicated in Fig. 11. Since $x_3 = x_2$, then the over-all differential equation for the system of Fig. 11 is

$$x_3{}^{(2)} + [a_1(t) - b_1(t)]x_3{}^{(1)} + [a_0(t) - b_0(t)]x_3 = x_1{}^{(2)} + a_1(t)x_1{}^{(1)} + a_0(t)x_1 \qquad (39)$$

Assuming that $x_1{}^{(2)}$, $x_1{}^{(1)} \to 0$ and x_1 as $t \to \infty$, theorem 1 in the section on stability can be applied to equation 39, which shows that

$$[a_1(t) - b_1(t)] = \bar{a}_1 - \bar{b}_1 = 2$$

$$[a_0(t) - b_0(t)] = 1 + e^{-2t} \to \bar{a}_0 - \bar{b}_0 = 1$$

$$a_1(t) = 2 + e^{-t} \to \bar{a}_1 = 2$$

$$a_0(t) = 2 + e^{-2t} \to \bar{a}_0 = 2 \qquad (40)$$

as $t \to \infty$; therefore

$$x_3(t) \to \bar{x}_3 = \overline{2x_1} \text{ as } t \to \infty \qquad (41)$$

and the system is stable.

To determine the sensitivity of the output x_3 to a small change in a_0, assume that a_0 is perturbed to $a_0 + \overline{\delta a_0}$ and the output is changed accordingly to $x_3 + \delta x_3$. Substituting these values into equation 39, the subsequent equation, in which second-order δ terms are neglected, is

$$\overline{\delta x_3}^{(2)} + [a_1(t) - b_1(t)]\overline{\delta x_3}^{(1)} + [a_0(t) - b_0(t)]\overline{\delta x_3} = \delta a_0[x_1 - x_3] \qquad (42)$$

Applying the conditions of relationships 40 and 41, the steady-state sensitivity of the system of Fig. 11 with respect to a differentially small change in a_0 is

$$E_{a_0}{}^T = \frac{\dfrac{\overline{\delta x_3}}{\bar{x}_3}}{\dfrac{\delta a_0}{\bar{a}_0}} = -\frac{\bar{a}_0 \bar{x}_1}{\bar{x}_3} = -1 \qquad (43)$$

Conclusions

The reduction techniques developed herein are no more general than the techniques for analyzing ordinary linear differential equations; their real advantage lies in the fact that the logical procedures of analysis, which have been developed for stationary signal flow graphs, are extended to nonstationary signal flow graphs in a much more useful way than that known heretofore.[4] An example of a situation in which the power of this technique is apparent is the stability analysis of a highly complex multiloop system. Using the reduction techniques developed in this paper, the signal flow graph can be rather easily reduced to an essential diagram. Each essential node can be examined individually to determine the over-all system stability. The individual node equations are of less order than that of the over-all system, thus in general reducing significantly the labor required in stability analysis. Also, the steady-state sensitivity of the over-all system is easily found from a knowledge of the differential equations of the individual transmittances, which is not true if weighting functions are used in the reduction process. This paper constitutes a fairly complete generalization of the concepts in chapter 2 of Truxal for the nonstationary case.[1]

References

1. Automatic Feedback Control System Synthesis (book), **John G. Truxal.** McGraw-Hill Book Company, Inc., New York, N. Y., 1955, chap. 2.

2. Properties of Impulsive Responses and Green's Functions, **K. S. Miller.** *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. CT-2, Mar. 1955, pp. 26–31.

3. On the Properties of the Impulsive Response Function of Systems with Variable Parameters, **V. Borskii.** *Automation and Remote Control*, vol. 20, July 1959, pp. 822–30.

4. On the Synthesis of Linear Automatic Control Systems with Variable Parameters, **S. V. Mal'chikov.** *Ibid.*, Dec. 1959, pp. 1543–49.

5. Linear Signal Transmission Systems, **L. A. Zadeh.** *Journal*, Franklin Institute, Philadelphia, Pa., vol. 253, Apr. 1952, pp. 293–312.

6. Random Processes in Automatic Control (book), **J. Halcombe Laning, Jr., Richard H. Battin.** McGraw-Hill Book Company, Inc., 1956, chap. 5.

7. Differential Equations: Geometric Theory (book), **S. Lefschetz.** Interscience Publishers, Inc., New York, N. Y., 1957, chap. 4.

8. Asymptotic Behavior and Stability Problems in Ordinary Differential Equations (book), **L. Cesari.** Julius Springer Verlag, Berlin, Germany, 1959, chap. 1.

9. Stability Theory of Differential Equations (book), **Richard Bellman.** McGraw-Hill Book Company, Inc., 1953, chap. 2.

10. Problems General de la Stabilite du Mouvement (book), **A. Liapunoff.** Princeton University Press, Princeton, N. J., 1949.

11. Nonstationary Signal Flow Graph Theory, **E. B. Stear, A. R. Stubberud.** *Report no. 60-64*, Department of Engineering, University of California, Los Angeles, Calif.

12. Non-stationary Smoothing and Prediction Using Network Theory Concepts, **S. Darlington.** *Transactions*, Professional Group on Information Theory, Institute of Radio Engineers, vol. IT-5, May 1959.

13. Ordinary Differential Equations (book), **E. L. Ince.** Dover Publications, Inc., New York, N. Y., 1956, chap. 11.

Automatic Plotting of Characteristic Curves and Analog Solution of Algebraic Equations

D. M. MITROVIĆ
NONMEMBER AIEE

Summary: For the characteristic equation $a_ns^n+a_{n-1}s^{n-1}+\ldots+a_2s^2+\xi s+\eta=0$ the relations between the coefficients ξ and η and the values $\zeta\omega_n$ and ω_n^2 are derived. The computer based on these relations may be used for plotting the characteristic curves used in the graphical analysis and synthesis of linear feedback control systems, for determination of roots of algebraic equations with real coefficients, and for plotting of root-locus for an arbitrary parameter.

IN PREVIOUS papers[1–5] a graphical method for analysis and synthesis of feedback control systems has been proposed by the author. The same method is also presented in reference 6. An essential role in this method is played by characteristic curves. Therefore, it is of interest to derive a method for automatic plotting of these curves.

As already shown,[1,6] the equations of the characteristic curves corresponding to the characteristic equation

$$f(s)=a_ns^n+a_{n-1}s^{n-1}+\ldots+a_2s^2+a_1s+a_0=0 \quad (1)$$

have the form

$$\xi=a_2\phi_2(\zeta)\omega_n+a_3\phi_3(\zeta)\omega_n^2+\ldots+a_n\phi_n(\zeta)\omega_n^{n-1} \quad (2)$$

$$\eta=-[a_2\phi_1(\zeta)\omega_n^2+a_3\phi_2(\zeta)\omega_n^3+\ldots+a_n\phi_{n-1}(\zeta)\omega_n^n]$$

where the functions $\phi_k(\zeta)$ are determined by the relation

$$\phi_k(\zeta)=-[2\zeta\phi_{k-1}(\zeta)+\phi_{k-2}(\zeta)] \quad (3)$$

with $\phi_0(\zeta)=0$ and $\phi_1(\zeta)=-1$.

The equations of 2 are obtained by introducing

$$s=-\zeta\omega_n+j\omega_n\sqrt{1-\zeta^2} \quad (4)$$

in equation 1 and considering a_1 and a_0 as the variables ξ and η. For a more detailed explanation of this development as well as for the table of functions $\phi_k(\zeta)$ see reference 1 or 6. As for ζ and ω_n, they represent, respectively, the relative damping coefficient and the undamped natural frequency. Both correspond to one pair of complex roots of the characteristic equation.

The equations of 2 may represent the loci of points in the $0\xi\eta$-plane corresponding to the roots with constant relative damping coefficient, with constant undamped natural frequency or with constant settling time depending on which variable among ζ, ω_n, and $\zeta\omega_n$ is considered constant.[1,2,6] The remaining variable plays the role of a variable parameter. In this paper, these characteristic curves will be called the damping curve, the natural frequency curve, and the settling time curve, respectively. The particular damping curves for $\zeta=0$ and $\zeta=1$ are called the stability curve and the critical damping curve. The automatic plotting of all these curves is the subject of the first part of this paper. The second part deals with the solution of algebraic equations based on the same principle.

Automatic Plotting of Characteristic Curves

Since the equations of 2 are not suitable for direct instrumentation, they shall be rearranged so that ξ and η become the functions of $\zeta\omega_n$ and ω_n^2 instead of ζ and ω_n. To this end, let us multiply equation 3 by ω_n^{k-1}. Thus,

$$\phi_k(\zeta)\omega_n^{k-1}=-[2\zeta\omega_n\phi_{k-1}(\zeta)\omega_n^{k-2}+\omega_n^2\phi_{k-2}(\zeta)\omega_n^{k-3}] \quad (5)$$

is obtained. As will be shown later, the expression $\phi_k(\zeta)\omega_n^{k-1}$ may be considered as the function of $\zeta\omega_n$ and ω_n^2 and denoted by $\psi_k(\zeta\omega_n, \omega_n^2)$, so that equation 5 becomes

$$\psi_k(\zeta\omega_n,\omega_n^2)=-[2\zeta\omega_n\psi_{k-1}(\zeta\omega_n,\omega_n^2)+\omega_n^2\psi_{k-2}(\zeta\omega_n,\omega_n^2)] \quad (6)$$

with $\psi_0=0$ and $\psi_1=-1$.

Since the starting values ψ_0 and ψ_1 are constants, the function ψ_2, because of equation 6, may be the function of $\zeta\omega_n$ and ω_n^2 only. The same may now be concluded for ψ_3 and so on. Thus, all ψ_k are the functions of $\zeta\omega_n$ and ω_n^2, which indicates that the denotation $\psi_k(\zeta\omega_n, \omega_n^2)$ is properly selected.

Paper 61-1053, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15–20, 1961. Manuscript submitted November 2, 1959; made available for printing September 6, 1961.

D. M. Mitrović is with the University of Belgrade, Belgrade, Yugoslavia.

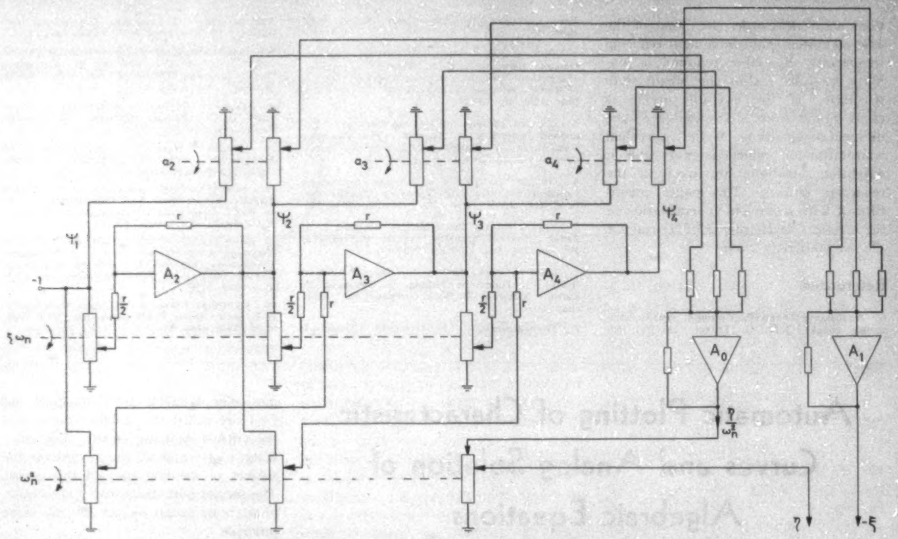


Fig. 1. Circuit diagram of the function generator for variables ξ and η

Because of equation 6, the basic equations of 2 may now be written in the required form

$$\xi = a_2\psi_2(\zeta\omega_n,\omega_n^2) + a_3\psi_3(\zeta\omega_n,\omega_n^2) + \ldots + a_n\psi_n(\zeta\omega_n,\omega_n^2)$$

$$\eta = -\omega_n^2[a_2\psi_1(\zeta\omega_n\ \omega_n^2) + a_3\psi_2(\zeta\omega_n,\omega_n^2) + \ldots + a_n\psi_{n-1}(\zeta\omega_n,\omega_n^2)] \quad (7)$$

in which ξ and η are expressed as functions of $\zeta\omega_n$ and ω_n^2.

Owing to equation 6, the equations of 7 are very suitable for instrumentation. The detailed circuit diagram of the function generator for ξ and η corresponding to a fourth order characteristic equation is shown in Fig. 1. The extension to higher order equations is evident.

The inputs $\zeta\omega_n$ and ω_n^2 are materialized as the shaft positions of the corresponding ganged potentiometers. Taking equation 6 into account, it is easy to see that the output of any amplifier A_k ($k=2,3,\ldots n$) presents the corresponding function $\psi_k(\zeta\omega_n,\omega_n^2)$. The remaining part of the circuit diagram performs the operations which are required by the equations of 7.

To plot a characteristic curve it is only necessary to apply the outputs $-\xi$ and η to a two-co-ordinate recorder and to vary one of the inputs, while keeping the other on a prescribed value. In this way, it is possible to plot directly either natural frequency curves (ω_n^2=constant) or settling time curves ($\zeta\omega_n$=constant). The plotting of the stability curve ($\zeta=0$) is performed by setting the shaft $\zeta\omega_n$ to its zero position and by varying ω_n^2. As for other damping curves (for $\zeta=\zeta_0\neq 0$), their plotting may be achieved only indirectly by calculating several pairs of $\zeta_0\omega_n$ and ω_n^2 for different values of ω_n and by introducing them as discrete input values. However, when two servomechanisms as in Fig. 2 are used, all damping curves can be directly plotted by setting the shaft ζ into the position corresponding to ζ_0 and by varying ω_n.

As evident from Fig. 1, any electronic differential analyzer with the drift compensated operational amplifiers may be used for automatic plotting of characteristic curves. However, it is also possible to construct a special completely transistorized computer of the a-c type, which is extremely desirable since the same device may be also used for finding the real and complex roots of algebraic equations.

Analog Solution of Algebraic Equations

Many methods have been proposed for analog solution of algebraic equations, but they all are really ineffective when solving for complex roots. However, in view of the method which shall be proposed here, this problem may now be considered successfully solved.

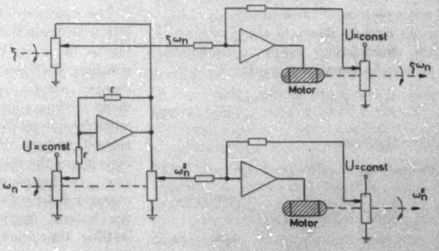


Fig. 2. Additional computations for direct plotting of damping curves

To explain this new method, let us start from equation 10 of reference 1:

$$f(-\zeta\omega_n+j\omega_n\sqrt{1-\zeta^2})=a_0-\eta+ (a_1-\xi)(-\zeta\omega_n+jw_n\sqrt{1-\zeta^2}) \quad (8)$$

which shows that $f(-\zeta\omega_n+j\omega_n\sqrt{1-\zeta^2})$ becomes zero, i.e., the given equation is satisfied only for those values of ζ and ω_n or $\zeta\omega_n$ and $\omega^2{}_n$ which determine ξ and η and satisfy the conditions $\xi=a_1$ and $\eta=a_0$.

Now, the use of analog computer for automatic plotting of characteristic curves is evident. It is only necessary to draw into the point $M(a_1, a_4)$ and, then, to plot the broken curve finishing at point M by the use of the computer. Positions of the shafts, $\zeta\omega_n$ and $\omega_n{}^2$, for which the recorder pen lies just in the point M (the conditions $\xi=a_1$ and $\eta=a_0$ are then satisfied) determine one pair of complex roots.

Instead of plotting the broken-curve by the recorder, the values ξ and η can be measured by using two voltmeters. The complex roots are then sought by adjusting the output variables ξ and η to the respective values a_1 and a_0. The positions of the shafts $\zeta\omega_n$ and $\omega_n{}^2$ for which equalities $\xi=a_1$ and $\eta=a_0$ are simultaneously satisfied determine one pair of complex roots. However, by using two additional potentiometers for setting the coefficients a_1 and a_0, the output variables ξ and η of the function generator can be transformed respectively into $\xi-a_1$ and $\eta-a_0$. The solution for complex roots is then reduced to simple adjustment of the new output variables to zero.

It is not difficult to see that the described computer can only be used for finding the complex roots with negative real part. By adding two amplifiers and a few switches (for signs of coefficients and for changes in circuit connections when solving for real roots), the computer can be made to find all the roots of any algebraic equation with real coefficients.

To determine the roots with absolute values larger than one, either the substitution $s=cv$ with $c>1$ or $s=1/v$ must be made.

Finally, it should be noted that the same procedure may be very conveniently programmed on a digital computer.

References

1. Graphical Analysis and Synthesis of Feedback Control Systems; I—Theory and Analysis, Dušan Mitrović. *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 77, 1958 (Jan. 1959 section), pp. 476–87.
2. Graphical Analysis and Synthesis of Feedback Control Systems; II—Synthesis, Dušan Mitrović. *Ibid.*, pp. 487–96.
3. Graphical Analysis and Synthesis of Feedback Control Systems; III—Sampled-data Feedback Control Systems, Dušan Mitrović. *Ibid.*, pp. 497–503.
4. Graphical Conditions for All Roots of an Algebraic Equation to Have Negative Real Parts (in French), D. M. Mitrović. *Comptes Rendus*, Academie des Sciences, Paris, France, vol. 240, 1955, pp. 1177–79.
5. Graphical Conditions for the Arguments of All Roots of an Algebraic Equation to Lie between $\pi/2+\mu$ and $3\pi/2-\mu$ (in French), D. M. Mitrović. *Ibid.*, vol. 243, 1956, pp. 831–33.
6. Analysis and Design of Feedback Control Systems (book), G. J. Thaler, R. G. Brown. McGraw-Hill Book Company, Inc., New York, N. Y., 1960, pp. 344–88.

◆

Discussion

George J. Thaler (U.S. Naval Postgraduate School, Monterey, Calif.): The analog methods suggested by Dr. Mitrović are very interesting, particularly since they permit plotting of any or all of the important characteristic curves. The only limiting factors seems to be the size of the analog facilities available, and for high order equations the use of a digital computer may be preferable. At the Naval Postgraduate School, we have prepared a digital program for our CDC (Control Data Corporation) *1604* computer which will compute the damping curves for equations up to order sixty-one. So far it has been used on equations up to the tenth order with satisfactory results. The program was based on equations 2 and 3 of the paper.

Another interesting aspect of Dr. Mitrović's analog scheme is that it appears possible to close the loop through servos and have the system seek the root values automatically. If this can be done, it may also be possible to incorporate design adjustments into the analog arrangement to permit experimental design for prespecified root locations.

D. M. Mitrović: I am obliged to Dr. Thaler for his valuable discussion which draws attention to the convenience of utilizing a digital computer. As for the suggestion of seeking the root values automatically, I think such a scheme is possible in the form of a two-dimensional nonlinear feedback system.

Analog Simulation of Digital Computer Programs

R. SAUCEDO
NONMEMBER AIEE

T. W. SZE
MEMBER AIEE

Summary: The objective of this paper is to reduce the concept of a digital computer program to an analogous sampled-data control system, which can then be studied and analyzed by established theories and simulated on an analog computer. In this connection, the convergence and the rate of convergence of various programs are established by observing the stability and the transient response on an analog computer.

THE USE OF digital computers as powerful tools in engineering, scientific, and business applications has been well established. During the past few years, digital computers began to appear as control elements in feedback control systems, first in air-borne equipments, then in petroleum-chemical, steel, and electrical utility industries. A principal difference of these control digital computers is that they operate on a "real-time" basis, and hence the speed of operation becomes a vital requirement for such operation.

While the speed of operation of a digital computer is primarily dependent on the speed of the computer components and the logical design of the computer, it is also heavily dependent on the digital program efficiency. Whether a digital program, which represents a numerical process, converges and how fast it converges to a required accuracy will materially influence the speed of operation.

The objectives of this paper are twofold. First, it will be attempted to reduce a digital program to an equivelent sampled-data control system for which the techniques of analysis are available.

Paper 61-1060, recommended by the AIEE Computing Devices Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Detroit, Mich., October 15-20, 1961. Manuscript submitted June 13, 1961; made available for printing September 11, 1961.

R. Saucedo is with the Westinghouse Electric Corporation, Pittsburgh, Pa., and T. W. Sze is with the University of Pittsburgh, Pittsburgh, Pa.

This work is based partially on a doctoral dissertation by R. Saucedo submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Pittsburgh.

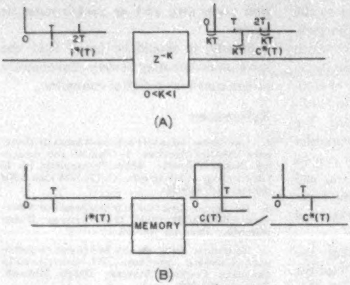


Fig. 1. Pure delay element behavior (A), operation of memory and nonsynchronous sampling element (B)

Second, it will be attempted to establish the analogs for various programs, so that various properties of digital programs such as the convergence and the rate of convergence may be studied by observing the stability performance and the transient response on an analog computer.

Digital Computer Programs and Analog Sampled-Data Systems

REALIZABILITY OF COMPUTER PROGRAM, GENERALIZED NECESSARY AND SUFFICIENT CONDITIONS

The general form of a digital computer program in the independent variable form may be expressed as

$$\sum_{k=0}^{n} f_k E^{-k} c_n - \sum_{k=0}^{n} g_k E^{-k} i_n = 0 \tag{1}$$

where the numerical operation indicated by $(E^{-k})c_n$ is defined by the following relation:

$$E^{-k} c_n = c_{n-k} \tag{2}$$

The f's and the g's in equation 1 need not be constants.

Equation 1 establishes a relation between the past and the present digital program inputs and the past and the present digital program outputs. The outputs and the inputs of the computer programs are defined as follows:

c_k = output at the kth sampling increment.
i_k = input at the kth sampling increment.
(3)

Consider the set D_1 which contains as its elements f_j for all j from 0 to n. Mathematically, this can be stated as follows:

$$f_j \epsilon D_1 \qquad j=0, 1, 2, \ldots, n \tag{4}$$

It can now be said that the digital program represented by equation 1 is physically realizable if the output depends only upon the present and past information of the inputs and only upon the past information of the output. Note that for this statement to hold, the following functional relations must apply:

$$f_j = f_j(i_0, i_1, \ldots, i_j;\ c_0, c_1, \ldots, c_{j-1}) \tag{5}$$

and

$$g_j = g_j(i_0, i_1, \ldots, i_j;\ c_0, c_1, \ldots, c_{j-1}) \tag{6}$$

In the above relations, it is assumed that a set D_2 has been defined in an analogous manner to D_1 so that

$$g_j \epsilon D_2 \qquad j=0, 1, 2, \ldots, n \tag{7}$$

Equation 1 may be rewritten and rearranged to yield

$$c_n = \frac{g_0 i_n + \sum_{k=1}^{n} E^{-k}(g_k i_n - f_k c_n)}{f_0} \tag{8}$$

Thus, if f_0, which is a quantity dependent on i_0, is unequal to zero, and if equations 5 and 6 are satisfied, the present information of the output is dependent upon

1. The present information of the input $g_0(i_0)i_n$
2. The past information of the input

$$\sum_{k=1}^{n} g_k(i_0, i_1, \ldots, i_k;\ c_0, c_1, \ldots, c_{k-1})E^{-k} i_n$$

3. The output

$$\sum_{k=1}^{n} f_k(i_0, i_1, \ldots, i_k;\ c_0, c_1, \ldots, c_{k-1})E^{-k} c_n$$

Note that if $f_0(i_0)=0$, then equation 1 becomes

$$f_1 c_{n-1} + \sum_{k=0}^{n} f_k E^{-k} c_n = \sum_{k=0}^{n} g_k E^{-k} i_n \tag{9}$$

$$c_{n-1} = \frac{g_0 i_n + g_1 i_{n-1} + \sum_{k=2}^{n} (g_k i_n - f_k c_n)E^{-k}}{f_1} \tag{10}$$

In a sequential operation such as in a digital program, equation 10 represents an impossibility, since c_{n-1} is dependent on i_n which is not available at the $(n-1)$th increment. To sum up, the necessary and sufficient conditions for physical realizability are:

$$f_j = f_j(i_0, i_1, i_2, \ldots, i_j;\ c_0, c_1, c_2, \ldots, c_{j-1})$$

$$g_j = g_j(i_0, i_1, i_2, \ldots, i_j;\ c_0, c_1, c_2, \ldots, c_{j-1})$$

$$f_0(i_0) \neq 0$$

PULSE TRANSFER FUNCTION OF A LINEAR COMPUTER PROGRAM

Let us consider a physically realizable linear computer program represented by

$$c_n = \frac{g_0 i_n + \sum_{k=1}^{n} E^{-k}(g_k i_n - f_k c_n)}{f_0} \tag{11}$$

where f_0 is unequal to zero and all f's and all g's are constants. Transforming equation 11 into the z-domain by the use of z-transformation, one writes

$$C(z) = \frac{g_0 I(z)}{f_0} + \frac{1}{f_0}\sum_{k=1}^{n} g_k Z^{-k} I(z) - \frac{1}{f_0}\sum_{k=1}^{n} f_k Z^{-k} C(z) \tag{12}$$

Defining the pulse transfer function $D(z)$ as the ratio of $C(z)$ to $I(z)$, one may write from equation 12 the following expression:

$$D(z) = \frac{C(z)}{I(z)} = \frac{\sum_{k=0}^{n} a_k z^{-k}}{1 + \sum_{k=1}^{n} b_k z^{-k}} \tag{13}$$

where

$a_k = g_k/f_0$ and $b_k = f_k/f_0$.

Equation 13 shows that a realizable linear computer program has a pulse transfer function expressed as the ratio of two polynomials in z^{-1}.

NONSYNCHRONIZED SAMPLED-DATA SYSTEM ANALOGS FOR LINEAR COMPUTER PROGRAMS

Equation 13 may be written as

$$D(z) = \frac{a_0}{1 + \sum_{k=1}^{n} b_k z^{-k}} + \frac{a_1 z^{-1}}{1 + \sum_{k=1}^{n} b_k z^{-k}} + \ldots + \frac{a_n z^{-n}}{1 + \sum_{k=1}^{n} b_k z^{-k}} \tag{14}$$

which is abbreviated as

$$D(z) = D_0(z) + D_1(z) + \ldots + D_n(z) \tag{15}$$

where

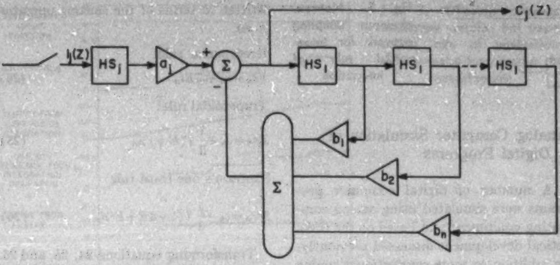


Fig. 2. Nonsynchronous sampling network

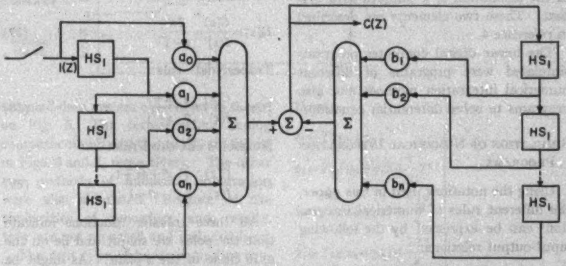


Fig. 3. Nonsynchronous sampling network for general pulse transfer function

$$D_j=\frac{a_j z^{-j}}{1+\sum_{k=1}^{n} b_k z^{-k}}=\frac{C_j(z)}{I_j(z)} \tag{16}$$

Equation 16 may then be rearranged as

$$C_j(z)=a_j I_j(z) z^{-j}-C_j(z)\sum_{k=1}^{n} b_k z^{-k} \tag{17}$$

Referring to the techniques of z-transformation in a sampled-data system,[1] the term z^{-1} pertains to a pure delay of T seconds, where T is the sampling period. Since it is quite difficult to simulate a pure delay on the analog computer, the pure delay elements are replaced by nonsynchronized sampling switches. The justification follows.

Fig. 1(A) shows the operation of a pure delay element. The duration of delay considered is KT seconds where $0<K<1$. Here the output signal is merely the input signal delayed by KT seconds, which is a fraction of the sampling period. The same input-output behavior may be achieved by the use of a memory element and a nonsynchronous sampler as shown in Fig. 1(B). Based upon this analog, the pure delay elements indicated by $z^{-\alpha}$ for any α can be represented. The memory and the nonsynchronous switch combination will henceforth be represented by $(HS)_\alpha$ in equations and in block diagrams, and the subscript will represent the delay in the switching time of the nonsynchronous sampler.

Thus, for the purpose of analog simulation, equation 17 is rewritten as

$$C_j(z)=a_j I_j(z)(HS)_j-C_j(z)\sum_{k=1}^{n} b_k (HS)_k \tag{18}$$

The sampled-data system realizing equation 18 is shown in Fig. 2. To represent a general computer program expressed by equation 14, Fig. 3 shows its equivalent sampled-data system that may be simulated by analog computer equipment.

NONSYNCHRONIZED SAMPLED-DATA SYSTEM ANALOGS FOR NONLINEAR COMPUTER PROGRAMS

A nonlinear computer program is one for which the f's and the g's in equation 11 are not constants. Without loss of any generality, consider the equation 11 with $f_0=1$:

$$c_n=\sum_{k=0}^{n} g_k E^{-k} i_n-\sum_{k=1}^{n} f_k E^{-k} c_n \tag{19}$$

Equation 19 shows that the present in-

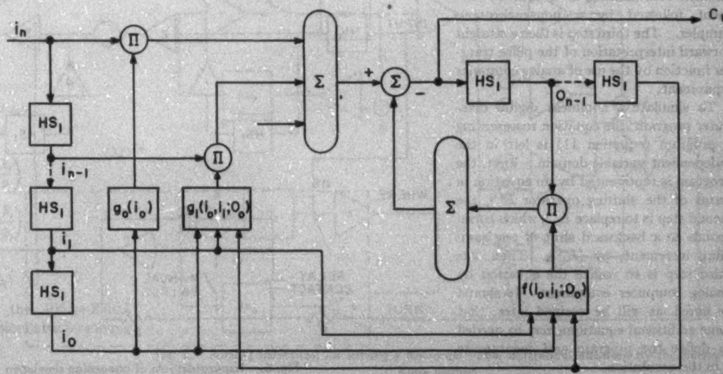


Fig. 4. Nonsynchronous sampling network for general nonlinear computer programs

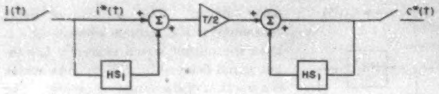


Fig. 5. Nonsynchronous sampling network for trapezoidal rule of integration

formation of the output is composed of the difference between two summations. The first summation deals with the past and the present information of the input properly weighted by the functions g_k. The second summation treats the past information of the output properly weighted by the functions f_k. Note that

$$E^{-j}c_n = c_{n-j}$$

and hence the output $E^{-j}c_n$ may be treated as c_n delayed by (j) seconds. Using this concept and the "*HS*" samplers to represent all shifting operators, equation 19 may now be written as

$$c_n = \sum_{k=0}^{n} g_k(HS)_k i_n - \sum_{k=1}^{n} f_k(HS)_k c_n \qquad (20)$$

Fig. 4 shows a sampled-data system simulation that realizes equation 20. Note that in the nonlinear computer program, g's and f's are not constants. Hence, to simulate such a computer program by analog computing equipment, the weighting factors g_k and f_k are no longer constants, and cannot be represented by constant potentiometer settings as multipliers.

The simulation of a digital computer program may be summarized in the following manner: To realize a linear digital computer program by a sampled-data system, it is useful first to obtain the pulse transfer function corresponding to the program. The second step is to replace z^{-1}, which corresponds to a pure delay element, by (*HS*) element, which corresponds to a tandem combination of a memory element followed by a nonsynchronous sampler. The third step is then a straight forward interpretation of the pulse transfer function by the use of analog computer equipment.

To simulate a nonlinear digital computer program, the equation representing a program (equation 11) is left in the independent variable domain. First, the program is represented by an equation in terms of the shifting operator E^{-1}, the second step is to replace E^{-1}, which corresponds to a backward shift of one sampling increment, by $(HS)_1$. Then, the third step is to realize the equation by analog computer equipment. It should be noted, as will be verified later, that some additional equations may be needed to define the program and incorporate it in the simulation.

Analog Computer Simulation of Digital Programs

A number of digital computer programs were simulated using analog computing equipment and based on the theoretical development discussed previously. In addition to some conventional analog computer components, two elements, the sampler and the holding circuit, were used in the simulation of a sampled-data system. These two elements are described in reference 4.

The linear digital computer programs simulated were programs of different numerical integration methods and also programs to solve differential equations.

SIMULATION OF NUMERICAL INTEGRATION PROGRAMS

Using the notations used in this paper, the different rules of numerical integration[3] can be expressed by the following input-output relations:

Rectangular rule:

$$c_{n+1} = c_n + Ti_{n+1} \qquad (21)$$

Trapezoidal rule:

$$c_{n+1} = c_n + \frac{1}{2}T(i_{n+1} + i_n) \qquad (22)$$

Simpson's one-third rule:

$$c_{n+2} = c_n + \frac{1}{3}T(i_n + 4i_{n+1} + i_{n+2}) \qquad (23)$$

These difference equations can be written in terms of the shifting operator E as:

Rectangular rule:

$$Ec_n = c_n + TEi_n \qquad (24)$$

Trapezoidal rule:

$$Ec_n = c_n + \frac{1}{2}T(E+1)i_n \qquad (25)$$

Simpson's one-third rule:

$$E^2c_n = c_n + \frac{1}{3}T(1+4E+E^2)i_n \qquad (26)$$

Transforming equations 24, 25, and 26 into z-domain, one may derive the pulse transfer functions for these programs:

Rectangular rule:

$$D(z) = \frac{C(z)}{I(z)} = \frac{Tz}{z-1} \qquad (27)$$

Trapezoidal rule:

$$D(z) = \frac{C(z)}{I(z)} = \frac{T}{2}\cdot\frac{z+1}{z-1} \qquad (28)$$

Simpson's one-third rule:

$$D(z) = \frac{C(z)}{I(z)} = \frac{T}{3}\cdot\frac{z^2+4z+1}{z^2-1} \qquad (29)$$

All three transfer functions indicate that the poles are simple and lie on the unit circle in the z-plane. As might be expected, one may conclude that all three numerical processes are convergent or stable.

The analog simulation of the programs may be carried out by first replacing a pure delay by a memory and a nonsynchronous sampler (*HS*). For example, equation 28 may be written as

$$C(z) = C(z)(HS)_1 + \frac{1}{2}T[1+(HS)_1]I(z) \qquad (30)$$

for the trapezoidal rule. The resulting

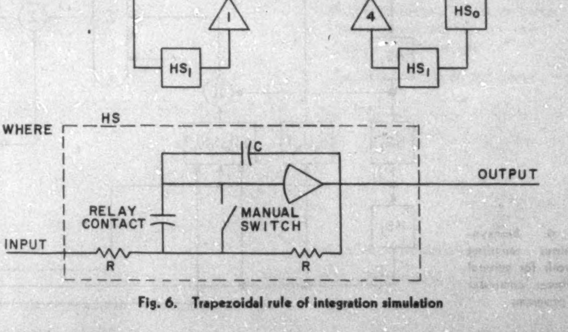


Fig. 6. Trapezoidal rule of integration simulation

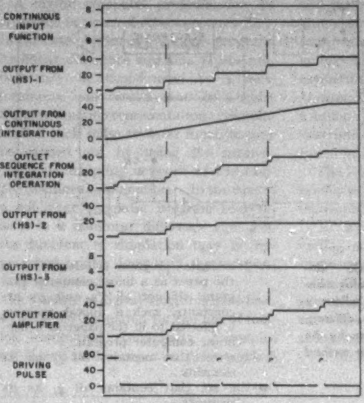


Fig. 7. Analog computer simulation results for trapezoidal rule of integration; see channel 4 from top for the output results

sampled-data system equivalent is shown on Fig. 5. The corresponding analog computer circuit and the results are shown in Figs. 6 and 7, respectively. The other two methods of numerical integration were also simulated. Because of the similarities of procedure and results, they are not repeated and included here.

The next type of linear computer program simulated is the Runge-Kutta method of solving differential equations.

SIMULATION OF RUNGE-KUTTA METHOD PROGRAM

Runge-Kutta method[2] as applied to a first-order differential equation may be stated as follows:

Given, that

$$\frac{dy}{dx}=f(x, y) \tag{31}$$

then the recurrent relations needed to solve the equation are

$$\left.\begin{aligned}
&x_{n+1}=x_n+T\\
&y_{n+1}=y_n+\Delta u_n\\
&\text{where}\\
&\Delta y_n=\frac{1}{6}(k_1+2k_2+2k_3+k_4)\\
&\quad k_1=Tf(x_n,\ y_n)\\
&\quad k_2=Tf\left(x_n+\frac{T}{2},\ y_n+\frac{k_1}{2}\right)\\
&\quad k_3=Tf\left(x_n+\frac{T}{2},\ y_n+\frac{k_2}{2}\right)\\
&\text{and}\\
&\quad k_4=Tf(x_n+T,\ y_n+k_3)
\end{aligned}\right\} \tag{32}$$

As an example, the Runge-Kutta method program was simulated for solving

$$\frac{dy}{dx}=x+y$$

Here,

$$\left.\begin{aligned}
&k_1=T(x_n+y_n)\\
&k_2=Tc_1(x_n+y_n)+\frac{1}{2}T^2\\
&k_3=Tc_2(x_n+y_n)+\frac{3T^2}{4}\\
&k_4=Tc_3(x_n+y_n)+\frac{7T^2}{4}\\
&\text{where}\\
&c_1=1+\frac{1}{2}T\\
&c_2=1+\frac{1}{2}T+\frac{T^2}{4}\\
&\text{and}\\
&c_3=1+T+\frac{1}{2}T^2+\frac{1}{4}T^3
\end{aligned}\right\} \tag{33}$$

The procedure and the method of simulation are the same as used for simulating the numerical integration previously. Figs. 8 and 9 show the analog computer circuit and the results respectively.

The simulation of the linear computer program just discussed is presented mainly to illustrate some of the simulation procedures. The next example is to illustrate the simulation of the nonlinear and iterative types of programs.

SIMULATION OF NEWTON-RAPHSON METHOD PROGRAM

The Newton-Raphson method[3] is a well-known numerical process used in solving for the solutions from a single equation or a system of equations that are nonlinear algebraic or nonlinear and transcendental. The method states that for an equation $f(x)=0$, the solution may be obtained by using the iterative formula

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \tag{34}$$

where $f'(x)$ is the first derivative of $f(x)$ with respect to x.

In order to illustrate the simulation procedure, a simple example using the Newton-Raphson method is chosen. It is a program of finding the reciprocal of a non-zero number b without using the division operation. Essentially, this program is to solve equation

$$f(x)=bx-1=0 \tag{35}$$

For this equation, the iterative formula of equation 34 becomes

$$x_{n+1}=x_n(2-bx_n) \tag{36}$$

Equation 36 may then be rewritten using the input-output notations as

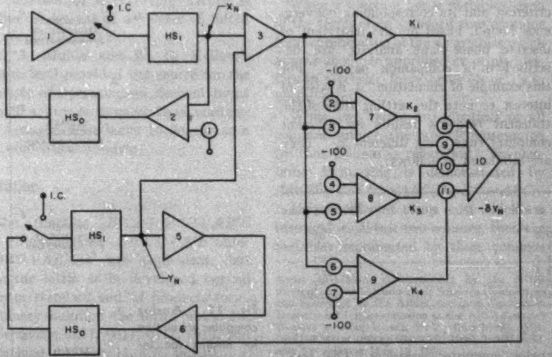


Fig. 8. Analog simulation for solving a first-order linear differential equation by the Runge-Kutta method

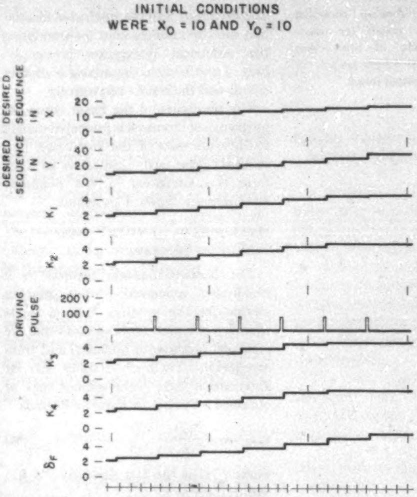


Fig. 9 (left). Results of analog simulation for solving a first-order differential equation by the Runge-Kutta method

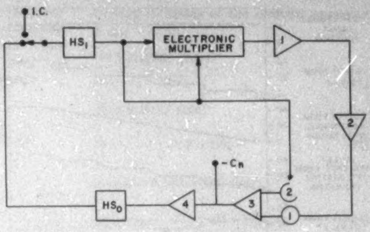


Fig. 10 (above). Analog simulation for finding a reciprocal by the Newton-Raphson method

$$c_n = i_n(2 - b i_n) \qquad (37)$$

The analog computer circuit is shown in Fig. 10, and the analog computer results are shown in Fig. 11. The cases studied were for $b = 0.5$ with initial approximations (x_o) used, spanning a range from 0.5 to 4.01. It was found that for x_o below 4.00, the equivalent analog system was stable. At $x_o = 4.01$ and beyond, computer results show that the equivalent analog system became unstable. On an analog basis, this means that the upper bound of convergence for this particular numerical process is 4.00. The theoretical investigation of such a stability criterion and its corresponding convergence bound, based on the treatment of discrete phase-plane analysis, are described in a companion paper. For this example of simulation, it is also of interest to note the settling time of the transient response results for different computer runs using different x_o. They are tabulated as follows:

x_0	Settling Time
0.5	5T
1.0	4T
1.5	2T
2.5	2T
3.0	3T
4.01	infinite

All these numbers are given within the analog computer accuracy. The comparison of settling time also shows the comparison of rate of convergence.

Nomenclature

E = a shifting operator, for example, $E^{-k}c_n = c_{n-k}$
c_n = output at the nth sampling increment
i_n = input at the nth sampling increment
f_h, g_k = the functional modifiers for the output and the input, respectively. If all f's and g's are constants, such a digital program is referred to in the paper as a linear computer program. If not all f's and g's are constants, such a digital program is referred to in the paper as a nonlinear computer program
D_1 = the set that contains all f_j as its elements
D_2 = the set that contains all g_j as its elements
(HS) = memory and nonsynchronous sampling switch combination to simulate a pure delay
z = variable in the z-domain
T = sampling period

Summary and Conclusions

The principal objective of this paper is to reduce the concept of a digital computer program to an analogous sampled-data control system, which can then be studied and analyzed by the sampled-data system theories and simulated on an analog computer.

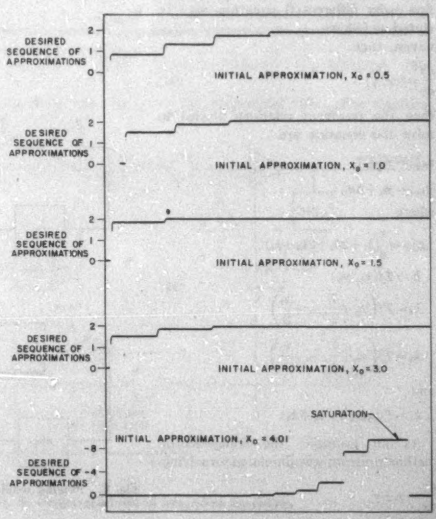


Fig. 11. Analog computer simulation for finding the inverse of 0.5 by the Newton - Raphson method

A digital computer program may be expressed in terms of difference equations. In the case of a linear computer program, its linear difference equation representation may be transformed by z-transformation and results then in a pulse transfer function representation. Simulation of such pulse transfer function may be carried out by using the nonsynchronous sampling network or the sampler-memory combination. In the case of a nonlinear computer program that results in a nonlinear difference equation, the problem of simulation may be approached in the domain of the independent variables.

A number of digital programs have been successfully simulated by the use of analog computing equipments. Examples are shown in this paper on the procedure and methods of simulation. Other programs simulated but not included in this paper are numerical prediction, linear homogeneous difference equations, roots extraction for polynomials using the Newton-Raphson method, and numerical solution of simultaneous equations. The description of these simulations may be found in reference 4.

The following are some of the possible applications for the concepts and methods developed in this paper:

1. Evaluation of a digital program by analog simulation to analyze the convergence and speed of convergence.
2. Comparison of relative efficiency of several digital programs that perform the same calculations but based on different numerical methods.
3. Simulation of real-time operation of a control system in which a control digital computer serves as a control element. In such simulation, the digital computer itself may be simulated by analog equipments, thus eliminating the necessity of coupling analog and digital computers together and the use of analog-to-digital and digital-to-analog equipments.

References

1. Digital and Sampled-data Control System (book), Joseph Tou. McGraw-Hill Book Company, Inc., New York, N. Y., 1959.
2. Digital Computer and Control Engineering (book), R. S. Ledley. McGraw-Hill Book Company, Inc., 1960.
3. Numerical Mathematical Analysis (book), J. B. Scarborough. The Johns Hopkins Press, Baltimore, Md., 1955.
4. Analysis of Some Digital Computer Programs by Use of Sampled-data Theory, R. Saucedo. *Ph.D. Thesis*, University of Pittsburgh, Pittsburgh, Pa., Jan. 1961.

Boolean Prime Implicants by the Binary Sieve Method

FRANK B. HALL
MEMBER AIEE

IN THE application of Boolean algebra to switching circuits,[1,2] one of the areas of interest is that of reduction of the functions to their simplest normal form. Quine, in formulating the special problem of simplification, proposed a method of determining what he called "prime implicants," and a method of selecting from the prime implicants the "simplest normal equivalents of a formula."[4] One of the requirements of this method is that the function first be expressed in "developed normal form" (standard sum of products or canonical form), each product of the function containing all of the letters of the function. Quine's later method eliminates this requirement.[5]

McCluskey, adopting the principles of Quine's first method, systematized the procedure generally, using binary notation instead of algebraic notation and introducing classification of the terms to reduce the number of comparisons required.[6] An algebraic method for generating all irredundant forms from the prime implicants is given by Petrick,[7] and in addition to these methods there is a voluminous literature in this area.

This paper deals with the first part of the problem of simplification as Quine proposed it, the determination of prime implicants. Selection among the prime implicants for the irredundant normal forms may be done according to the methods cited, and is beyond the scope of this paper.

With this method, as in Quine's first method, the function must first be expressed as a standard sum of products, each term containing every variable. The prime implicants are disclosed through the action of a hypothetical binary sieve. Computation of a recurring nature is eliminated by use of a "Table of Sifted Numbers," Table I.

A. A. Mullin and W. G. Kellner's residue test[8] provided the source for the principle of operating on decimal terms in such a way as to cause those containing a single redundancy to appear as a pair of identical numbers.

Notation

The function $B'CD+A'CD+AB'C$ and the function $AB'CD+A'B'CD+A'BCD+AB'CD'$ are equivalent, but only the latter is in developed normal form or standard sum of products form. In binary notation the terms would be, respectively, 1011, 0011, 0111, 1010, and in decimal notation, 11, 3, 7, and 10.

In binary notation each column represents a power of two just as in the decimal system each column represents a power of ten. The columns represent $2^0, 2^1, 2^2, 2^3, 2^4 \ldots$, or the equivalents 1, 2, 4, 8, 16 . . ., but with the order reversed. It is very convenient to designate a column as the "4 column" or the "8 column," and for a given function each column is associated with a particular variable. In this paper "column" is used in the sense of "bit position in a binary number."

Example

64	32	16	8	4	2	1	
2^6	2^5	2^4	2^3	2^2	2^1	2^0	
1	0	1	0	0	1	1	binary notation
A	B'	C	D'	E'	F	G	algebraic
64	+	16		+	2 +	1 = 83	decimal

Redundant Variables

When a function has n variables, each of the terms of its standard sum contains some combination of the n variables which appear as either unprimed or primed letters in algebraic form, or as 0's or 1's in binary form.

When two terms are identical in all columns but one, the variable represented by that column is redundant. In algebraic form this is demonstrated by $ABCD+ABCD'=ABC(D+D')=ABC$.

If four different terms exist which are identical in all but two columns, then the variables represented by those columns

---

Paper 60-1283-A, recommended by the AIEE Computing Devices and Basic Sciences Committees and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted April 29, 1960; made available for printing May 19, 1961.

Frank B. Hall is with the Particle Accelerator Division (PAD), Argonne National Laboratory, Argonne, Ill.

are redundant and can be eliminated from those terms. If eight different terms exist which are identical in all but three columns, then the variables represented by those columns are redundant and can be eliminated from those terms.

The general rule is that if 2^m different terms exist which are identical in all but m columns, then the variables represented by those columns are redundant and can be eliminated from those terms.[9]

Prime Implicants by the Binary Sieve Method

The principle of this method can best be explained by imagining a sieve capable of sifting out the 1's in any given column of binary numbers and leaving 0's in their places. Suppose such a sieve were to sift out the 1's in the 2 column. The number 11000 (decimal form 24) having a zero in that column would drop through without change. The number 11010 (decimal form 26) would lose a 1 from its 2 column and drop through as 11000 (decimal form 24).

Suppose that the sieve is to sift out the 1's in the k column, and suppose that a number of terms in binary form are dropped into the sieve. Those terms with a 1 in the k column will lose it and change. Those that have a 0 in the k column will be unchanged. For each number that has changed in passing through the sieve, there are now two numbers identical in all but the k column: the number before sifting and the number that resulted from the sifting.

If all the numbers were to be listed after each sifting, then two numbers that differed only in the k column would appear in the k sifting as a pair of identical numbers; the one with a 0 in the k column would be unchanged and the one with a 1 in the k column would lose it, thus changing into the other. If a multiple sifting is used, that is, if the terms are sifted for one column and the results immediately sifted again for one or more other columns, then the general rule for redundant variables can be applied directly. Four identical terms resulting from two siftings reveal two redundancies, eight identical terms from three siftings reveal three redundancies, etc. For manual computation, however, this kind of multiple sifting is impractical. A start is made by sifting each column once, after which the pairs of identical terms revealed in this way are used to generate larger combinations in a manner that is equivalent to multiple siftings.

Given a function in standard sum form, its terms are written in a row designated for convenience as the "T row," followed by the 1, 2, 4, 8, . . . 2^n rows. The entries in the 1 row result when the T row terms are sifted to remove the 1's in the 1 column. The entries in the 2 row result when the T row terms are sifted to remove the 1's in the 2 column. This is continued until there is a row corresponding to each column of the variables of the function.

In each case an entry is made only if the term in the T row has changed in passing through the sieve. If any sifted terms also appear in the T row, that is, if any of the sifted terms are terms of the function, then these terms are encircled (shown boldface in this paper because of type limitations) and are designated as "essential sifted terms." The result of this procedure will be a sifted terms array.

Example

T row:	0001	0010	0100	1000	1011	1100	1101	1111
1 row:	0000	——	——	——	1010	——	**1100**	1110
2 row:	——	0000	——	——	1001	——	——	**1101**
4 row:	——	——	0000	——	——	**1100**	1001	**1011**
8 row:	——	——	——	0000	0011	**0100**	0101	0111

In decimal form the principle of the array is the same, but the sifted terms are copied from Table I. The table has been constructed in the same way as the array above, but the terms have been shown in decimal rather than binary form. The same array in decimal form is shown below.

Example

T row:	1	2	4	8	11	12	13	15
1 row:	0	–	–	–	10	—	**12**	14
2 row:	–	0	–	–	9	—	—	**13**
4 row:	–	–	0	–	—	**8**	9	**11**
8 row:	–	–	–	0	3	**4**	5	7

It is now obvious from the construction of the array and the definition of essential sifted term that each encircled (boldface) term and the T row term from which it was sifted are identical in all but one column. The number of that column corresponds to the number of the row. Each pair reveals a redundant variable, allowing the two terms to be combined into one with that variable eliminated.

In order to effect larger combinations of terms in this way, two general theorems are now given which state the conditions for finding additional redundant variables.

Theorem 1: In a sifted terms array, if 2^j different terms identical in all but j columns can be found in the k row, where the k column is not one of the j columns, then with the corresponding terms in the T row they form 2^{j+1} different terms identical in all but $j+1$ columns.

Proof: In a sifted terms array, a sifted term appears in the k row only if the corresponding term in the T row has changed when sifted for the k column. That is, it appears in the k row only if the number from which it was sifted has a 1 in the k column. Obviously, a sifted term in the k row has a 0 in the k column. In no case, therefore, will any of the 2^j terms in the T row duplicate any of the 2^j terms in the k row.

However, in all columns except the k column, each T term is identical to its corresponding sifted term in the k row. It is given that the 2^j terms in the k row are different terms identical in all but j columns, and that the k column is not one of the j columns. It follows that the 2^j terms in the T row must form *another* set of 2^j different terms identical in all but j columns. It has now been shown that there are 2^{j+1} different terms and that these terms are identical in all but j and the k columns, or in $j+1$ columns.

Theorem 2: In a sifted terms array, if 2^j different terms in the T row, identical in all but j columns, have 2^j sifted terms in the k row, where the k column is not one of the j columns, then with these sifted terms they form 2^{j+1} different terms identical in all but $j+1$ terms.

The proof of this theorem is almost identical to that for Theorem 1, except that it is now shown that the 2^j sifted terms must form another set of 2^j different terms identical in all but j columns. Otherwise reasoning is the same.

It should be noted that the theorems apply to all sifted terms, some of which without further restriction might not be terms of the given function. This difficulty is easily overcome by simply ignoring all oridinary sifted terms and considering only essential sifted terms. This way no term can enter a combination unless it is a terms of the function.

Each time one of these theorems is used, the effect is to eliminate another variable and to double the number of terms in a combination. Any encircled term and its associated T row term form a pair to start with.

Assume that the standard sum of a given function consists of the following terms: 0, 1, 3, 8, 9, 13, 14, 15, 16, 17, 19, 24, 25, 27, 31. These steps are taken:

1. Write the terms in sequence as the T row of an array.
2. Under each T row write the sifted terms as found in Table I.
3. Encircle those sifted terms that also

Table I. Table of Sifted Numbers

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
(1)		0		2		4		6		8		10		12		14
(2)			0	1			4	5			8	9			12	13
(4)					0	1	2	3					8	9	10	11
(8)									0	1	2	3	4	5	6	7
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
(1)		16		18		20		22		24		26		28		30
(2)			16	17			20	21			24	25			28	29
(4)					16	17	18	19					24	25	26	27
(8)									16	17	18	19	20	21	22	23
(16)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	32	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**45**
(1)		32		34		36		38		40		42		44		47
(2)			32	33			36	37			40	41			44	46
(4)					32	33	34	35					40	41	42	43
(8)									32	33	34	35	36	37	38	39
(16)																
(32)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	48	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**
(1)		48		50		52		54		56		58		60		62
(2)			48	49			52	53			56	57			60	61
(4)					48	49	50	51					56	57	58	59
(8)									48	49	50	51	52	53	54	55
(16)	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
(32)	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	64	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**
(1)		64		66		68		70		72		74		76		78
(2)			64	65			68	69			72	73			76	77
(4)					64	65	66	67					72	73	74	75
(8)									64	65	66	67	68	69	70	71
(16)																
(32)																
(64)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	80	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**95**
(1)		80		82		84		86		88		90		92		94
(2)			80	81			84	85			88	89			92	93
(4)					80	81	82	83					88	89	90	91
(8)									80	81	82	83	84	85	86	87
(16)	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
(32)																
(64)	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	96	**97**	**98**	**99**	**100**	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**	**111**
(1)		96		98		100		102		104		106		108		110
(2)			96	97			100	101			104	105			108	109
(4)					96	97	98	99					104	105	106	107
(8)									96	97	98	99	100	101	102	103
(16)																
(32)	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
(64)	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
	112	**113**	**114**	**115**	**116**	**117**	**118**	**119**	**120**	**121**	**122**	**123**	**124**	**125**	**126**	**127**
(1)		112		114		116		118		120		122		124		126
(2)			112	113			116	117			120	121			124	125
(4)					112	113	114	115					120	121	122	123
(8)									112	113	114	115	116	117	118	119
(16)	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
(32)	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
(64)	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63

appear in the *T* row (shown boldface).

4. Ignore entirely any unencircled sifted terms in the array.
5. Write each essential sifted term and its associated *T* row term as a pair, and group them by row number. Those from the *j* row will be referred to as "*j* sifting."

The left-hand members of these siftings are essential sifted terms from various rows. The right-hand members are associated terms from the *T* row.

Beginning with the first pair in the 1 sifting, each combination of terms in that sifting is compared with the left-hand members of the 2, 4, 8, 16 siftings. Similarly, the combinations in the 2 sifting are compared with the 4, 8, and 16 siftings, the 4-sifting combinations with the 8 and 16 siftings, and the 8 sifting with the 16 sifting. The combinations in the 16 sifting are examined only to see whether or not they have appeared in one of the larger combinations generated by the previous comparisons.

T row:	0	1	3	8	9	13	14	15	16	17	19	24	25	27	31
1 row:	–	0	2	–	8	12	–	14	–	16	18	–	24	26	30
2 row:	–	–	1	–	–	–	12	13	–	–	17	–	–	25	29
4 row:	–	–	–	–	–	9	10	11	–	–	–	–	–	–	27
8 row:	–	–	–	0	1	5	6	7	–	–	–	16	17	19	23
16 row:	–	–	–	–	–	–	–	–	0	1	3	8	9	11	15

Siftings

(1)		(2)		(4)		(8)		(16)	
	0–1		1–3		9–13		0–8		0–16
	8–9		13–15		27–31		1–9		1–17
	14–15		17–19				16–24		3–19
	16–17		25–27				17–25		8–24
	24–25						19–27		9–25
									15–31

Each time the terms of a combination are found in the left-hand members of another sifting, Theorem 1 is applicable. The associated right-hand members of that sifting also become members of the combination.

Taking 0-1 as an initial combination, one looks, unsuccessfully in this case, for 0 and 1 in the left-hand members of the 2 and the 4 siftings. However, they are found in the 8 sifting paired with 8 and 9. The terms of the new combination 0-1-8-9 are then found in the left-hand members of the 16-sifting paired with 16, 17, 24, 25. Inasmuch as there are no further siftings to check, the combination 0-1-8-9-16-17-24-25 is taken as the final combination. As a matter of convenience, the numbers of the siftings that are involved in this combination are also listed, thus (1, 8, 16).

Every possible combination starting with each pair in the siftings must be tested in this way with Theorem 1. Once a combination has been generated, part of that combination is used as a start in testing for other combinations. Terms are dropped off as they were added, but in inverse order, those added last being dropped off first, and each time terms are dropped a new attempt is made to generate another combination with the terms remaining. This proceeds until the only terms remaining are the initial pair and until that pair generates no new combination.

If a combination is generated, all of whose terms are already members of some other combination, the smaller combination is discharged.

From the combination generated above, 0-1-8-9 is selected as a start to find a new combination, but is quickly found to be unsuccessful. A start is now made with 0-1 which generates only 0-1-16-17, but this must be discarded because it is only part of the original combination. Check off 0-1.

The 8-9 combination generates nothing but (1, 16) 8-9-24-25, all of whose terms already appear in the first combination of (1, 8, 16) 0-1-8-9-16-17-24-25. The smaller combination is therefore discarded, and the 8-9 combination is then checked off.

There is a useful observation that shortens the work when it is applicable. If the initial pair being tested appears in some combination already generated, it will generate no new combinations unless the terms also appear in some sifting other than those associated with that combination. In the preceding paragraph, for instance, 8 and 9 are part of a combination with 1, 8, 16 as associated sifting numbers. If 8 and 9 do not appear in at least one of the remaining siftings (2 sifting or 4 sifting), the pair may be checked off at once without further testing. Obviously the usefulness of this device grows with the presence of large combinations.

The terms of the next combination in the 1 sifting, 14-15, are found in no other sifting, nor are they included by any previously generated combination. The pair 14-15 is therefore listed as a final combination.

This procedure continues, progressively shortening as each sifting is exhausted. The combinations in the 16 sifting are tested only to see if they are already included by some combination. If so, they are checked off. If not, they are listed as final combinations.

Finally, one must list among these final combinations any terms in the T row that do not have essential sifted terms. (No such term appears in the given example.) When carried out completely, this procedure generates the following combinations:

(1, 8, 16)	0–1–8–9–16–17–24–25
(1)	14–15
(2, 16)	1–3–17–19
(2)	13–15
(2, 8)	17–19–25–27
(4)	9–13
(4)	27–31
(16)	15–31

Failure to test the combinations exhaustively as outlined in the procedure given may result in missing some larger combination. A very useful checking technique is to take each final combination into the right-hand members of the siftings, using Theorem 2. Should all the terms of some combination appear there, the left-hand members join the combination. If this should occur, it will usually happen that the new combination will include other lesser combinations, allowing them to be discarded.

These combinations, after conversion to algebraic form, represent the prime implicants, and the process of selecting among them for the simplest normal equivalents is the remaining part of the minimization treated by Quine, McCluskey, and Petrick.[4,6,7] It is beyond the scope of this paper.

Conversion of such combinations to algebraic form is now simple.[10] (a) Take the smallest number of the combination. (b) Write it in binary form. (c) Substitute dashes for the zeros in the columns indicated by the sifting numbers. (d) Write the algebraic form from the remaining zeros and ones.

Example

(1, 8, 16) 0–1–8–9–16–17–24–25
(a) 0 (b) 0 0 0 0 0 (c) – – 0 0 – (d) $C'D'$

(2, 8) 17–19–25–27
(a) 17 (b) 1 0 0 0 1 (c) 1 – 0 – 1 (d) $AC'E$

Short Version of the Method

Some familiarity with the method will suggest the possibility of eliminating the explicit listings of the siftings. It is entirely practical to work directly from the array. Each encircled term and its associated T row term form the initial combination which is then looked for among the encircled terms of subsequent rows. The associated T row terms then join the combination under the conditions of Theorem 1. The process of checking off the terms is identical; a light dash drawn through the encircled terms is a convenient mark.

Construction of a Sifted Terms Array Without Use of Table

If a table of sifted numbers is not at hand, the array may easily be formed by developing the sifted terms by computation. This is preferably done in two steps: (1) Indicate by dots the positions of sifted terms. (2) Calculate the sifted terms by subtracting the row number from the T row term.

If the binary forms of the T row terms were to be written vertically, the position of the 1's would correspond to the occurrence of the sifted terms. Hence it is possible to place the dots by inspection.

Another method is to mark all the sifted term positions in one row before proceeding to the next. Each T row term is divided by the row number (1, 2, 4, ...) and if the quotient is odd a sifted term is indicated and a dot marks the position. Any fraction in the division is neglected. The theorem upon which this is based is due to Mullin and Kellner.[8]

Example

Given 4, 5, 8, 10, 11, 15, form a sifted array.

Step 1.

	4	5	8	10	11	15
(1)		·			·	·
(2)				·	·	·
(4)	·	·				·
(8)			·	·	·	·

Step 2.

	4	5	8	10	11	15
(1)	–	4	–	–	10	14
(2)	–	–	–	8	9	13
(4)	0	1	–	–	–	11
(8)	–	–	0	2	3	7

Summary

This paper describes a method of determining the prime implicants of Boolean functions in the standard sum of products form. It is somewhat akin to the Quine–McCluskey method, but differs considerably in procedure, being based on relations disclosed by the action of a hypothetical sieve capable of sifting out 1's from any given binary column of a set of numbers. The method is specifically adapted for manual computation with the decimal form of notation. Although the Table of Sifted Numbers facilitates the procedure, simple computation can be substituted when the table is not available.

Appendix

That the method of this paper will produce all the prime implicants, and only the prime implicants, of a function can be demonstrated. Rephrasing somewhat the definitions given by Quine,[4] a term is said to subsume another term if all of the variables of the latter term are found among the variables of the first term, primed or unprimed in the same manner. A term is said to be a "prime implicant" of a function if an only if it implies the function and subsumes no shorter term implying the function.

Assume a function of n variables in expanded normal form with at least one prime implicant composed of r of the n variables, where $r<n$. By repeated applications of the Boolean algebra theorem that $A=A(B+B')=AB+AB'$, it can be shown that the existence of such a prime implicant is equivalent to the existence of 2^{n-r} different terms identical in all but $n-r$ columns. With respect to one of the nonidentical columns, say k, the 2^{n-r} terms can always be divided into two combinations of 2^{n-r-1} different terms such that the terms of one contain all 1's in the k column and the terms of the other combination contain all 0's in the k column.

Each combination will then constitute 2^{n-r-1} different terms identical in all but $n-r-1$ columns, one combination of terms occurring in the k row of an array of the function as essential sifted terms of the other combination of terms. The terms in the k row resulting from the first division can then be divided similarly with respect to one of the remaining nonidentical columns, say the h column. Each of these new combinations will have 2^{n-r-2} terms, and one combination would appear in the h row of an array as essential sifted terms of the other. This "reduction" in terms will eventually end when but two terms remain, one of which will be an essential sifted term of the other. These would necessarily appear as a pair in one of the siftings from the array. At each step of this "reduction" process, the inverse operation of combination by Theorem 1 is obviously possible. The effect of this reasoning is to show that any prime implicant in which less than all the variables of a function appear is equivalent to a combination of terms which *can* be generated by Theorem 1 from some pair in one of the siftings of an array of the function. Inasmuch as the procedure in this method demands that Theorem 1 be applied repeatedly to every such pair in the siftings until it fails to generate a combination, it must generate every combination to which it is applicable. It, therefore, must generate the combination equivalent of the prime implicant assumed above, and hence those of all prime implicants.

An important part of the procedure of this method is the discarding of combinations which are part of some larger combination. To show that this method produces only prime implicants, the significance of this is explored. A combination of 2^{n-r-1} different terms identical in all but $n-r-1$ columns is equivalent to a term with $r+1$ variables. A combination of 2^{n-r} different terms identical in all but $n-r$ columns is equivalent to a term with r variables. If the 2^{n-r-1} terms are part of the larger combination of 2^{n-r} terms, they are necessarily identical in all but $n-r$ columns. The term with the $r+1$ variables then subsumes the shorter term with the r variables. Discarding combinations that are part of some larger combination of terms is therefore equivalent to discarding terms that subsume shorter terms. Because all prime implicants are generated, all the shortest terms implying the function are displayed. The terms that survive the discarding subsume no shorter terms implying the function, and are therefore prime implicants. Hence only prime implicants are produced.

Expansion of a Function

If the function contains terms having less than all the variables, it is not in standard sum or canonical form, and each such term must be expanded into terms containing all variables. In decimal form it is quite easy to expand such terms in the following manner: (a) Write the term in binary form, indicating the missing variables by dashes. (b) Write the term again, substituting zeroes for the dashes. (c) In parentheses write the decimal equivalents of the missing variables, and beside them the decimal equivalent of the term. (d) Finally, take all the sums obtained by adding the numbers in parentheses to the original number in all possible combinations.

Example

Assume function has five variables, A, B, C, D, E.

$AB'E$ (a) 10 - - 1 (b) 10001 (c) (2,4) 17
(d) 17, 19, 21, 23

AC (a) 1-1 - - (b) 10100 (c) (1,2,8) 20
(d) 20, 21, 22, 23, 28, 29, 30, 31

References

1. A Symbolic Analysis of Relay and Switching Circuits, **Claude E. Shannon.** *AIEE Transactions*, vol. 57, 1938, pp. 713–23.

2. The Design of Switching Circuits (book), **W. Keister, A. Ritchie, S. Washburn.** D. Van Nostrand Company, Inc., Princeton, N. J., second edition, 1952.

3. Switching Circuits and Logical Design (book), **S. H. Caldwell.** John Wiley & Sons, Inc., New York, N. Y., 1958.

4. A Problem of Simplifying Truth Functions, **W. V. Quine.** *American Mathematical Monthly*, Menasha, Wis., vol. 59, Oct. 1952, pp. 521–31.

5. A Way to Simplify Truth Functions, **W. V. Quine.** *Ibid.*, vol. 62, Nov. 1955, pp. 627–31.

6. Minimization of Boolean Functions, **E. McCluskey, Jr.** *Bell System Technical Journal*, New York, N. Y., vol. 35, Nov. 1956, pp. 1417–44.

7. A Direct Determination of the Irredundant Forms of a Boolean Function from the Set of Prime Implicants, **S. R. Petrick.** *AFCRC-TR-56-110*, Air Force Cambridge Research Center, Bedford, Mass., Apr. 1956.

8. A Residue Test for Boolean Functions. *American Mathematical Monthly*, vol. 59, Oct. 1952, pp. 162–64.

9. See reference 2, p. 91.

10. See reference 3, pp. 168–69.

X-Ray Tube with Grid for Control of Exposures in Cineradiography

T. H. ROGERS
MEMBER AIEE

CINERADIOGRAPHY of the human body has long provided valuable diagnostic information, not otherwise obtainable, in cases where motion of anatomical structures is involved. This method is particularly valuable in detecting functional or structural abnormalities in the heart organ, and also in examining the larynx, gastro-intestinal tract, joints, and other moving parts.

The earliest method of obtaining a series of individual X-ray pictures required for cine-presentation was to photograph with a motion-picture camera a continuously produced visible image on a fluoroscopic screen. Exposure of patients to X rays was necessarily restricted. Thus, the X-ray dosage rate required to produce an image bright enough for successful motion-picture photography imposed severe limitations on the duration, and consequently the usefulness, of cinematic studies made in this manner.

Advances in X-ray image intensification by electronic means have greatly reduced the X-ray intensity required to produce sufficiently bright images, so that cineradiography is now a widely used diagnostic procedure. However, further refinements in techniques and apparatus are being constantly sought and evolved, with the twofold object of increasing diagnostic effectiveness and decreasing further the X-ray dosage needed.[1]

All elements in the cineradiographic process have received attention in the search for improvement—the X-ray generator, image intensifier, optical system, camera, film, film development, projector, and the interdependence of all these elements. Theoretical investigations, based on quantum physics, have shown that there is an absolute minimum to the amount of X-radiation that will produce an image containing a required level of information, no matter how effective the image intensifier may be in producing increased brightness, how efficient the optical system, how fast the film.

Improvement in the over-all system may be obtained by two basic approaches: (1) minimizing the number of quanta lost, wasted, or otherwise not used in producing the images actually photographed; and (2) minimizing loss of information in the process of recording and projecting the images.

Conventional Exposure Control System

A major reduction in wasted quanta has been achieved by arranging the X-ray apparatus so that X rays are not generated while the camera shutter is closed

---

Paper 61-34, recommended by the AIEE Electrical Techniques in Medicine and Biology Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 29–February 3, 1961. Manuscript submitted February 25, 1960; made available for printing November 18, 1960.

T. H. Rogers is with Machlett Laboratories, Inc., Springdale, Conn.

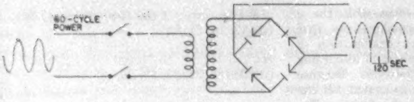


Fig. 1 (above). Schematic diagram of conventional X-ray generator with primary switching

during the film pull-down interval. This is normally accomplished by operating the exposure control switch of the X-ray machine in synchronism with the camera shutter to turn the high voltage on and off for each "frame." This requires, of course, that timing and switching devices in the exposure control mechanism repeat their operations many times per second for long periods with high accuracy. The conventional exposure control actuates a contactor supplying 60-cycle power to the primary of the high-tension transformer. Usually the transformer output is full-wave rectified, so that the voltage waveform applied to the X-ray tube is as illustrated schematically in Fig. 1.

The number of half-cycles used per exposure may depend on the number of cine-frames per second to be recorded. A frequently employed system provides optionally 7.5, 15, 30, or 60 frames per second, with X-ray time settings of 1/15, 1/30, 1/60, or 1/120 second, respectively. These times, of course, represent 8, 4, 2, or 1 half-cycles, respectively, of rectified voltage to the X-ray tube. It is also possible, with such a system, to employ a single half-cycle, or 1/120-second exposure for each frame, regardless of the frame rate, provided it is a submultiple of 60 per second and the exposure is properly synchronized.

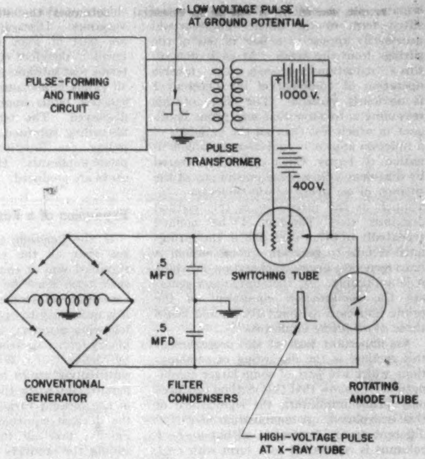


Fig. 2 (right). Schematic diagram of circuit for electronic switching in high-voltage secondary

High-Voltage Switching

The very short exposure represented by the single half-cycle reduces image blurring due to motion. A lower patient dose would be necessary to produce the required image if the voltage pulse producing the X radiation could have a square shape, rather than the half-sine-wave shape delivered by the 60-cycle transformer-rectifier. Hence, systems have been developed to switch the output voltage of the high-voltage generator to the X-ray tube by means of a high-voltage electronic switch tube.[3] The switch-on interval may be made short enough to include only a small portion of the half-sine-wave and synchronized to take only the peak portion thereof. If a large filter-capacitor is also employed, such synchronization may not be necessary. Fig. 2 shows schematically the basic circuit for such a system.

Systems of this type have been successfully employed, but their limitations became increasingly significant as further improvements in the techniques of image intensification and fluorography were evolved. Generally speaking, the voltage pulse must be fed to the X-ray tube through a long, shockproof cable which has a fairly large inherent capacitance. The tube does not begin to generate X rays at full voltage until the cable capacitance has been completely charged. And after the voltage pulse is switched off, the cable discharges through the X-ray tube, continuing to produce X rays at decreasing voltage levels during the discharge interval.

Today's efficient intensifiers require relatively low tube currents, even for millisecond (ms) exposures. Also, there is an increasing tendency to employ long runs of shockproof cable. The combination of large capacitance with small discharge-current value results in a great stretch-out of the X-ray pulse, with a long "tail" of low-energy components. Therefore, much of the benefit of the initial short, square voltage pulse is lost. Fig. 3 shows tracings of oscillograms of the current waveform obtained with 30-foot cables, with tube currents ranging from 100 ma (milliamperes) to 600 ma, at 80 kv.

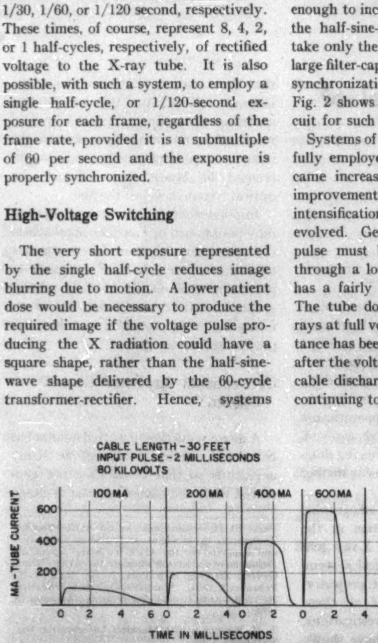


Fig. 3 (left). Current-pulse shapes affected by cable capacitance

Grid-Controlled X-Ray Tube

The obvious way to overcome the described loss is to interrupt the current at a point immediately adjacent to the X-ray tube, so that cable-capacitance charging and discharging does not occur during the exposure period. To use the high-voltage switching tube in this manner would entail placing a bulky piece of apparatus in an inconvenient location,

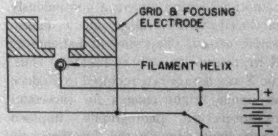


Fig. 4 (right). Schematic of grid-control and beam-forming structure—transverse section

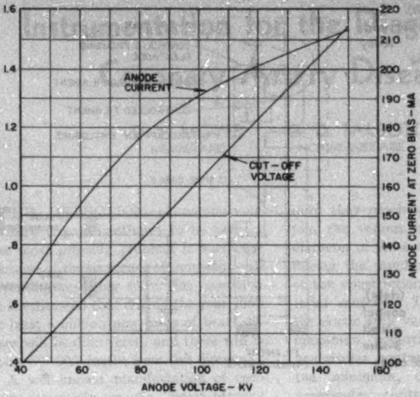


Fig. 5. Cutoff-grid voltage and anode-current characteristics of grid-controlled tube

physically near to the X-ray tube. A preferable possible alternative was to use a control grid in the X-ray tube itself.

Grid-controlled X-ray tubes were employed in the past when trying to control exposure duration and other parameters such as anode current and focal-spot size. Generally speaking, such attempts were not commercially successful because of the apparent impossibility of achieving proper control of any one parameter without undesirably affecting others. The requirements for a structure that permits cutoff with a desirably low value of negative grid = voltage are incompatible with a structure designed to produce suitably high anode currents and to control their values while maintaining a proper focal spot. At this juncture, re-examination of X-ray tube grid control problems—specifically in the light of present-day cineradiography requirements—was undertaken.

As far as control is concerned, the X-ray-tube current, voltage, and pulse duration per exposure need to be determined. A careful analysis of all factors involved indicates that suitable characteristics would include a maximum of 150-kv X-ray-tube voltage, a maximum of 100-ma peak-tube current, and individual frame exposure, or pulse duration, ranging from 0.5 to 4.0 ms. Pulse-repetition rates up to 60 or more per second, and total sequence duration up to 100 seconds, might be required.

The design of the cathode and grid system, then, can be "boiled down" to a structure for simple bistable operation, that is, for complete cutoff with a manageable value of negative-grid voltage and for full-emission current when the negative-grid potential is removed. The full-emission current will be determined primarily by the filament temperature as in conventional X-ray tubes, rather than by grid-driving voltage. The grid would be at cathode potential, or at some fixed potential close thereto, during the exposure period. Thus, the focal-spot size and configuration can be controlled by conventional means, independently of anode current. The absence of any positive-grid voltage requirement eliminates grid current and its attendant complications in the control circuit.

Analytical and experimental exploration of cathode-grid design geometry promptly indicated and confirmed the necessity for a compromise between maximum anode current and minimum cutoff-grid voltage in a zero-grid-drive design. Such compromise is a less serious matter, of course, if the required peak-anode current is relatively small. The problem is still further minimized if the focal-spot size can be made smaller than normal. All these considerations led to a quite simple design in which the cathode-focusing member and the grid were combined into a single structure.

When in the CUTOFF condition, this structure acts as a grid to suppress the electron emission from the filament. When ON, the same structure focuses the emitted electrons into a suitable cross-sectional beam to produce the desired focal spot on the anode. Careful optimization of the structure's dimensional details, by combined analysis and experimental trial and error, has produced a set of characteristics consistent with previously established needs. Fig. 4 shows a transverse cross section through the structure, indicating how cutoff occurs when the grid is negative and how beam-forming is accomplished when the grid is at filament potential. The focal spot is of the conventional line-focus type.

Grid-cutoff voltage as a function of anode voltage and anode current as a function of anode voltage at zero-grid voltage with a fixed allowable value of filament temperature are charted in Fig. 5.

A number of parallel problems, requiring solution in connection with incorporating the grid-controlled cathode structure in a tube to perform fully the required functions, will be described briefly in subsequent paragraphs.

Considerations leading to the choice of a focal spot approximately 0.5 mm (millimeter) wide were, primarily, that the image detail suffers with an appreciably larger focus and, secondarily, that such a small size reduces the compromise between minimum cutoff voltage and maximum anode current. However, a basic limitation is target overheating in the focal area when maximum current and voltage are utilized simultaneously. The well-known rotating-anode principle is, *a priori*, employed to reduce such heating, but at the maximum desired-peak loading of 100 ma at 150 kv, a 0.5-mm focal spot on the conventional rotating target is instantaneously overloaded.

To insure latitude in the loading capacity for such demands, provisions for possible higher rotation speeds, up to 10,000 rpm (or 3 times normal), have been developed and incorporated into the grid-controlled tube. Inasmuch as instantaneous loading capacity is, by classic theory, proportional to the square root of the rotation speed, the maximum permissible focal-spot loading is almost doubled. The range of maximum individual exposure factors available under various conditions of operation is indicated in Table I.

Practical usage of a tube having the described characteristics was determined by consulting numerous potential customers, who invariably pointed out that in any cineradiography installation, the ability to perform conventional radiog-

Table I. Allowable Loadings for Square Pulses of 4 Ms or Less

Voltage Kv Peak	Current Ma Peak	
	Normal Rotation	Triple Rotation
100	100	170
125	80	140
150	67	115

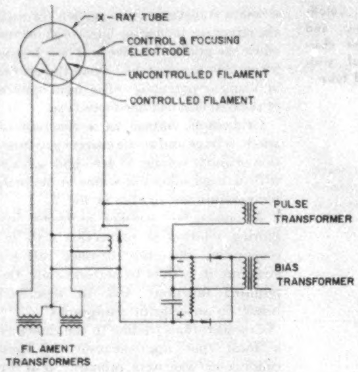


Fig. 6 (left). Control circuit for grid-controlled X-ray tube

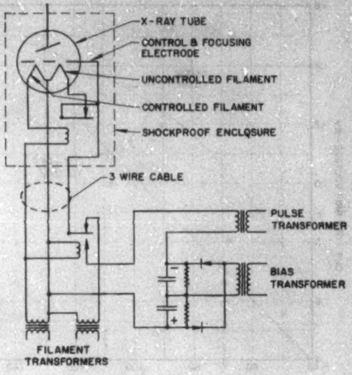


Fig. 7 (right). Modified control circuit to permit use of 3-wire shockproof cable

raphy, and particularly spot-film radiography, would be desired. Most spot-film techniques require X-ray-tube currents as high as 300 to 500 ma. To provide loading capacity at this level would necessitate a much larger focal spot. Hence, in addition to the desired small focus with grid control for cine work, a second cathode, minus this control and able to produce a larger focal spot for conventional high-current radiography, was provided in the same tube. The second focus is the nominal 1.5-mm size, suitable for 300-ma techniques at normal rotational speed and 500 ma at triple speed.

Control Circuitry

The circuitry involved in applying filament-power and control-grid voltage to such a tube becomes somewhat complicated. A basic circuit for this purpose, diagrammed in Fig. 6, has the bias voltage removed to produce anode-current pulses by a cancellation-voltage pulse in the grid circuit.

Retention of the standard 3-wire cable with the present standard cable terminal and receptacle for feeding supply voltages to the grid-controlled tube would be desirable. However, the control-grid voltage sources must be incorporated in the X-ray-tube head if no more than three conductors are to be employed in the shockproof cable feeding all required voltages for the proposed arrangement.

The needed components can be reduced to a reasonably compact package but, when added to the usual tube-head dimensions, would create an undesirably large unit for use under an X-ray table. Two arrangements, calculated to minimize or eliminate this disadvantage, were devised. Either scheme involves a small auxiliary unit or "black box," which can be mounted in the table and connected between the X-ray generator and tube by acting as a junction box in the negative shockproof cable.

One arrangement, illustrated schematically in Fig. 7, retains the 3-wire cable standard and involves a relay enclosed in a small appendage to the X-ray-tube head. The other, employing the circuit shown in Fig. 6, involves no additional elements in the tube head but does require a 4-wire cable between the black box and the tube. In either case, the interlead capacitance of the connecting cable should be kept as low as possible (and the cable as short as possible) to minimize the charging current, which the pulse transformer must supply. If this current drain is low, a very simple pulse transformer is feasible.

The comparative desirability of the two approaches has not been established but is under exploration. Both are being employed currently in systems installed for clinical use and evaluation. Fig. 8 shows the pulse shape of X-ray intensity, obtained when operating under pulsing conditions at 60 pulses per second, 1-ms pulse duration, 80 kv, 60-ma-peak-tube current. The oscillogram was taken of a scintillometer output exposed to the output of the pulsed X-ray tube.

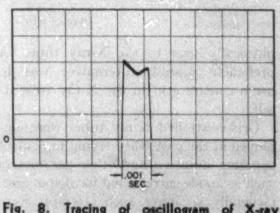


Fig. 8. Tracing of oscillogram of X-ray output from grid-controlled tube, 1-ms pulse duration, as detected by scintillometer

Conclusions

A tube designed to produce X-ray exposures of optimum characteristics for cineradiography with today's X-ray image-intensifier systems is now available, and is being incorporated into X-ray installations which are suitably designed to take advantage of these characteristics. Reports on clinical experience at such installations may be expected soon.

References

1. CINEFLUOROGRAPHY. *Proceedings*, First Annual Symposium on Cinefluorography. Charles C Thomas, Springfield, Ill., 1960.

2. HIGH TENSION SWITCH TUBE FOR ROENTGENOGRAPHY, Charles T. Dotter, T. H. Rogers. *The American Journal of Roentgenology, Radium Therapy and Nuclear Medicine*, Charles C Thomas, vol. LXXV, Jan. 1956, pp. 83–90.

Instrumentation for the Diagnosis of Coronary-Artery Disease

D. F. MOYER
MEMBER AIEE

G. D. TALBOTT
NONMEMBER AIEE

THE DETECTION of coronary-artery disease continues to be perhaps the most pressing problem in medicine, because of the incidence of morbidity and mortality resulting from this condition. It is estimated that during the year 1959, at least 4 million new cases of heart disease will be discovered, and there will be some 500,000 deaths from this disease.[1]

A well-known manifestation of coronary-artery disease is a "heart attack," in which an artery supplying blood to the heart muscle is suddenly occluded. This constantly working muscle, when deprived of its nourishment, immediately signals its distress by sharp chest pain. The severity of the attack depends on the extent and rapidity of the disruption of blood supply; but all too often, this first indication of coronary-artery disease terminates rapidly in death.

Detection of coronary-artery disease by electro-acoustical means is currently under investigation by the authors, and this paper deals with some early results of this work. It is hoped that these results will lead to a quantitative evaluation of the disease, and even to the localization of a specific portion of the artery where plaques are present and an occlusion may be threatening. This evaluation will not only indicate the magnitude of the danger, but will monitor the results of therapeutic efforts and measure their effectiveness.

Heart Dynamics

The coronary arterial system is made up of two major arteries originating at the aortic arch, which is the output tube for oxygenated blood from the heart. These arteries then branch and subdivide, constantly decreasing in diameter until they reach capillary size, there to join the venous system. The effective diameter of these arteries changes in size during the regular heart cycle, because of the constriction of the heart and arterial muscles. The blood pressure in the aortic arch varies between systolic—maximum, occurring immediately after ventricular contraction, and diastolic—the minimum, occurring just before ventricular contraction. As a result, the blood flow through the coronary arteries occurs in an uneven but repetitive cycle.

A sound is generated in this passage of blood through the coronary arteries that is propagated as pressure variations through the tissues and air cavities of the body. The pressure level of this arterial sound is considerably below that of the heartbeat heard through a stethoscope. The familiar "lub-dup" of the beating heart is the back-pressure wave generated largely by the closing of the heart valves, and may easily be heard on the chest wall.[2] Because of the low level of the coronary-artery sounds, it is fortunate that most (66%) of the coronary blood flow occurs during diastole, or when the heart is at rest.[3]

During this high coronary flow period, the sounds of the heart filling with blood and those of respiration are also present. Respiration noises can be practically eliminated by having the patient hold his breath, but separation of the sounds of the heart filling must be ob-

tained b
obvious t
normal h
was foun
artery di
distributi
As coron
there is
artery.
and, at l
restricts
off. Th
interior v
occurs in
longitudi
cause of
section c
siderably
is believ
in the bl
eration p
in sound
diseased

Cadaver

It was
condition
coronary
the sou
those so
on cadav
the soun
up in th
it was u
obtained
hours af
came fr
18-80, r
the wh
acoustic
opened l
degree c

METHO

There
where t

Fig. 1. Left-descending coronary artery with plaques at points (A)

Paper 60-100, recommended by the AIEE Electrical Techniques in Medicine and Biology Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., January 31–February 5, 1960. Manuscript submitted November 2, 1959; made available for printing December 17, 1959.

D. F. MOYER, Consulting Engineer, and G. D. TALBOTT, M.D., are of Dayton, Ohio.

The authors are indebted to Dr. James Funkhouser for the pathological evaluation of the coronary arteries, and to Dr. Melvin Oosting and his staff for their co-operation.

Fig. 2. Cadaverous heart under test

supplied from the aorta. These arteries are the right (posterior) descending artery and the left (anterior) artery which splits into a descending (toward the bottom, or apex, of the heart) and circumflex artery (which circles the heart at its largest diameter). A catheter was inserted into each of these ostia in turn, and saline solution under 120 mm Hg (millimeters of mercury) pressure was admitted. Because of the congestion of blood in the fine capillary ends of the arteries, there was little, if any, flow established; but the arterial system swelled under pressure and was readily discernible. An incision was made at a capillary section of the artery under pressure, and flow was established. It should be noted that this method allows the saline solution to simulate the action of blood only through that section of the arterial network which is drained by the incision. The most vital areas of sclerosis are in the largest sections of the arteries, and the incision, which is usually well downstream from these sections, should cause a flow simulating the normal flow through them. The rate of flow in cubic centimeters/minutes under 120 mm Hg was measured by the timed filling of a graduate.

METHOD OF MEASUREMENT

Once the flow conditions were established, a microphone probe was placed against the heart muscle near a large section of the artery under test; see Fig. 2. By means of a d-c solenoid valve, a pulse of saline was applied through the catheter into the coronary artery.[4] A tape recorder with a tape loop running about 8 seconds was used to record the output of the microphone. Two or three pulses of about $1^{1}/_{2}$ seconds' duration were recorded, along with a clear tape section of no saline flow. This latter section of tape was recorded to establish the over-all noise signal of the system. The analysis of the signal is always accomplished in conjunction with analysis of the noise. This allows for use of recordings having the very low signal-to-noise ratio common in this low-level work.

Recordings were also made of patients with a previous history of coronary-artery disease. These were obtained using the same microphone without the probe extension. The probe was placed on the chest wall immediately above the heart, and the patient suspended breathing during the test.

Instrumentation

The microphone used in this work is a diaphragm, variable-reluctance type: see Fig. 3. Affixed to the center of the 0.008-inch steel diaphragm, 2 inches in diameter (F) or (K), is a brass probe (N) about 3/4-inch long and 0.130 inch in diameter. This probe is threaded to receive a brass extension used in the cadaverous heart work, and to reach the heart directly through chest-wall incisions in animals. A soft, sponge-rubber cup (B) is attached to the steel case (A) of the microphone. This cup shields the diaphragm from air-born noise while the microphone is pressed firmly against the chest of the patient, and serves to distribute the pressure evenly. When the necessary pressure is applied, the rubber compresses and allows the probe to firmly contact the chest wall. It is interesting to note that the patient cannot feel the high per-unit pressure of the small metal probe because of the rubber ring; even though the mark of the probe is clearly visible on the skin after removal.

A permanent magnet (P) sets up the necessary flux in the two pole pieces (E) and diaphragm. A coil (G) surrounds each pole, and the coils are connected in series additive. As the probe responds to pressure variations on the surface of the skin, it moves the diaphragm, causing a reluctance change in the magnetic circuit. This reluctance change varies the flux ϕ in the pole pieces (E) according to the relation $\phi = F/R$, where F is the magnetomotive force of the permanent magnet (P) and R is the magnetic circuit reluctance. A signal voltage e is generated in the coils (G) by the changing flux according to the relation $e = N d\phi/dt$, where N is the number of additive coil turns and $d\phi/dt$ is the rate of change of flux. A steel plate (M) about 0.094-inch thick is spaced by a brass ring (L) about 0.050 inch in front of the diaphragm. This, together with the steel case, serves as an acoustical and magnetic shield for the coils. The plate has a concentric, cylindrical tube around the probe which is filled with a high-viscosity grease (Q) to serve as an acoustical absorbant, preventing air-born noise from reaching the diaphragm. A Bakelite shell (D) is used as an insulator for mounting the components, and the assembly is held

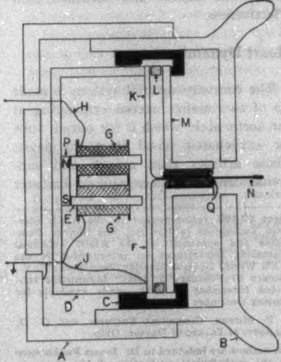


Fig. 3. Magnetic microphone

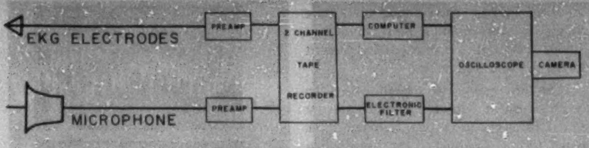


Fig. 4. Measuring system

together by the Bakelite cover (*C*) screwed onto the shell (*D*).

The low impedance (about 1,000 ohms) of these coils allows a long lead to the preamplifier where an input transformer applies the signal to the grid of a shock-mounted *6112* dual triode. The preamplifier has no a-c in its enclosure (the filament is rectified a-c) to keep down 60-cycle pickup. Although the microphone coils are well shielded, and are of low impedance, they are subject to electromagnetic pickup such as is found in hospital facilities.

The preamplifier drives a tape recorder through its associated a-c amplifier, making a permanent record on the tape loop; see Fig. 4. This tape is then suitably identified with a tag carrying the medical information of the subject, and brought to the laboratory for analysis. The tape is played through a variable band pass electronic filter. The filter is of the twin *T* negative feedback type with separately adjustable high-pass and low-pass sections in series. The output of the filter appears on an oscilloscope, where a picture is made for a permanent record.

A marker generator is used to select a portion of the signal for repeated analysis. The mark is put on a second channel of the same tape, so that it will not interfere with the signal, but will provide a repeatable reference point. The tape loop is then cycled and the scope's sweep is triggered by the marker. The sound is analyzed while examining a series of one-octave bands by means of the electronic filter.

Results

A representative analysis of four hearts is presented here. Fig. 5 shows oscillograms of a cadaverous heart having minimal sclerosis. Another cadaverous heart having advanced coronary-artery disease is illustrated by oscillograms shown in Fig. 6. Analysis of a white male, aged 70, with a history of rheumatic heart disease but no clinical symptoms of sclerosis, is pictured in Fig. 7. The last set, Fig. 8, is from a 65-year-old, white male, having a history of severe coronary-artery disease and a prior coronary occlusion.

Fig. 9(B), which is also shown in Fig. 6(C), is the signal recorded from the microphone on the heart while being pulsed with a saline solution. Fig. 9(A) is the identical condition a moment later, when the saline flow has stopped (a companion analysis of the over-all noise). The illuminated grids on all of the oscillograms are 1 cm (centimeter) square, and the vertical and horizontal deflections are given in millivolts (mv) and milliseconds (ms) per centimeter, respectively.

First, contrast the sclerotic and nonsclerotic cadaverous hearts in Figs. 5 and 6. Except for relative amplitude (based on the over-all or unfiltered signal) Figs. 5(D) and 6(D) are similar. Both oscillograms show roughly the same predominate frequency, which, although it is not a pure sine wave, has its peaks fairly evenly spaced along the time axis. Note, however, that the main frequency component for the sclerotic heart has an amplitude modulation or "pulsed" characteristic.

In the 300–600 BPF (band-pass filter) setting, the relative amplitudes of the two cadaverous hearts remain unchanged (Fig. 5 has increased vertical gain) from the 150–300 BPF oscillogram. Note that the predominate frequency component is considerably higher, and both sclerotic and nonsclerotic hearts show evidence of pulsing.

The 600–1,200 BPF oscillograms of Figs. 5 and 6 show that the relative amplitude of the nonsclerotic heart sounds diminish and the pulsing largely disappears. The amplitude has diminished from about 1,200-mv peak-to-peak (pp) in the 150–300 BPF, to 250 mv pp in the 600–1,200 band-pass filter. The sclerotic heart, while showing a diminution of the pulsing phenomena, is still maintaining

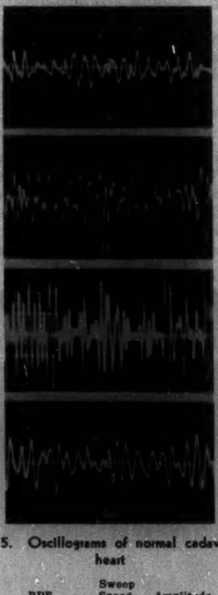


Fig. 5. Oscillograms of normal cadaverous heart

BPF, Cps	Sweep Speed, Ms/Div	Amplitude, Mv/Div
1,200–2,500	2	50
600–1,200	5	100
300–600	10	200
150–300	10	500

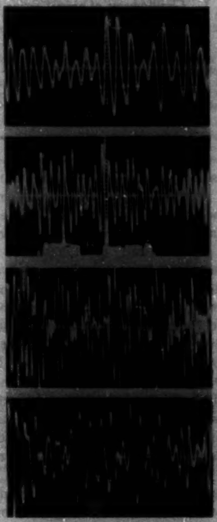


Fig. 6. Oscillograms of sclerotic cadaverous heart

BPF, Cps	Sweep Speed, Ms/Div	Amplitude, Mv/Div
1,200–2,500	2	100
600–1,200	5	500
300–600	10	400
150–300	10	500

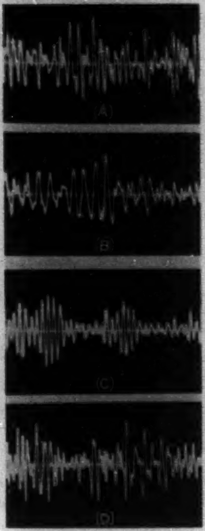


Fig. 7. Oscillograms of normal active heart

BPF, Cps	Sweep Speed, Ms/Div	Amplitude, Mv/Div
1,200-2,500	2	5
600-1,200	5	20
300-600	10	100
150-300	20	100

the same amplitude as in the 150–300 BPF (vertical sensitivity 500 mv/cm).

The most pronounced difference between the two recordings occurs when the BPF is adjusted to 1,200-2,500 cps (cycles per second). Here, the nonsclerotic heart (Fig. 5) has minimal pulsing, and its amplitude is reduced to about 75 mv pp. The sclerotic heart, however, has an amplitude of about 500 mv pp. Fig. 10 shows the same 1,200–2,500 BPF setting at a slower sweep speed, 5 ms/cm to show how the pulsing persists. It should be noted that the relative amplitudes of the various oscillograms may also be dependent on factors other than sclerosis.—Note that the peaks of the signal occur rather regularly in the top oscillogram of Fig. 6, about 1 ms apart—indicating a frequency of around 1 kc. This frequency occurs outside the pass band of a filter which falls off 24 decibels/octave. A further check with the BPF set to 1,000–1,100 cps in Fig. 11 disclosed a virtually unchanged picture of larger amplitude, 750 mv pp of the 1-kc frequency. Other sclerotic hearts show this amplitude modulated fundamental in the range of 1,000–1,500 cps. It was hardly present in the nonsclerotic hearts.

It is interesting to speculate on the cause of this modulated, high-frequency sound in sclerotic coronary arteries. A paper recently published by Bruns on a general theory of the cause of cardiac murmurs advances the idea of periodic detachment of vortices as the producer of sound.[5] As stated earlier in this paper, the growth of a plaque will form a constriction in the arterial tube which will be followed by a segment of larger cross-section. This abrupt increase in effective diameter will cause vortices to form on the downstream side. As these vortices detach, they could give rise to the amplitude modulation of a fundamental caused by the vortices themselves. An experimental proof of this hypothesis has not been attempted by the authors.

The purpose for investigating the cadaverous hearts was to determine an approximate coronary sound field with a minimum amount of unwanted signal. As noted earlier, the maximum coronary flow occurs during diastole. The ventricles of the heart are also filling at this time, but if there are no valvular abnormalities, there is little sound generated. If there is mitral or tricuspid stenosis (the valves between the auricle and ventricle do not fully open), a sound, usually on the order of 100 cps, is generated. In the apparently normal heart, therefore, the sound caused by coronary flow should be a maximum during diastole, while that caused by other heart action should be a minimum.

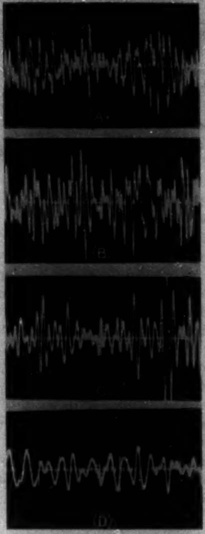


Fig. 8. Oscillograms of sclerotic active heart

BPF, Cps	Sweep Speed, Ms/Div	Amplitude, Mv/Div
1,200-2,500	2	2
600-1,200	5	2
300-600	10	10
150-300	10	50

To study the heart signal in human subjects during diastole, an electrocardiogram (EKG) signal was obtained on a separate tape channel simultaneously with the microphone signal. The QRS complex of the EKG signal (which occurs at the beginning of each heartbeat) was used to trigger a transistorized flip-flop circuit, to give a sharp marking pulse on the second channel. This marking pulse was used to trigger an adjustable time delay on the oscilloscope sweep generator. The oscillogram in Fig. 12, shows the normal heartbeat of the 70-year-old, white male, with the microphone in the proximity of the aortic area. The horizontal sweep is 200 ms/cm and the brightened portion begins about 500 ms after the marking pulse has triggered the sweep. The oscilloscope actually has two time-base generators, one of which triggers the other. The sweep rate of both generators is separately adjustable, and the starting point of the second generator is continuously adjustable anywhere along the 10-cm sweep of the first generator. In Fig. 12, the first generator is sweeping at 200 ms/cm, and the second generator is triggered at 2.5 cm or 500 ms. The second generator is set to sweep at 50

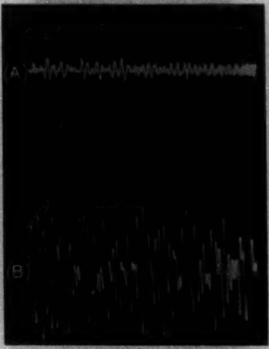


Fig. 9. Signal/noise oscillogram

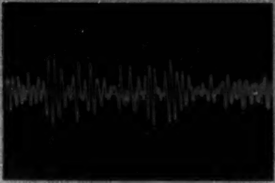


Fig. 10. Slower oscillogram of sclerotic cadaverous heart

Fig. 11. Narrow-band oscillogram of sclerotic cadaverous heart

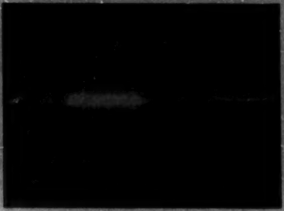


Fig. 12. Heartbeat oscillogram—brightened portion showing diastolic phase

ms/cm and, thus, would require 500 ms for covering the full 10 cm of the scope. Five-hundred ms at a rate of 200 ms/cm would occupy 2.5 cm; thus, the brightened portion of the trace is 2.5 cm long in Fig. 12. The brightened portion is, thus, set to show only the diastolic sounds, and the repetitive analysis of Figs. 7 and 8 was accomplished from the QRS marking pulse in this manner. The delay time from the beginning of the trace to the beginning of the brightened portion was held constant in Figs. 7 and 8; these series show only the brightened portion at various sweep rates. The sweep speeds and BPF settings were maintained consistent with the cadaverous heart analysis (except for the lower picture of Fig. 7 being 20 ms/cm instead of 10), but the vertical sensitivity was necessarily increased to show the fainter sounds appearing at the chest wall.

Note that the patient with valvular involvement caused by rheumatic fever has marked pulsing in the 300–600 cps range, but progressively less for the higher filter ranges. On the other hand, the patient with symptomatic coronary-artery disease has less evident pulsing in the lower ranges, and more evident pulsing in the higher BPF frequencies. Use of the octave filter frequencies described in this paper gives an indication of the pulsing in a general way. A more detailed analysis of the sound spectrum, using a more narrow frequency band, gives a clearer indication of the modulation phenomena, as shown in oscillogram of Fig. 11. Investigation is continuing along this line to identify more precisely the sounds generated by blood flow in coronary arteries with plaque formation

Conclusions

The investigation described in this paper was instigated to develop instrumentation for the early diagnosis of coronary-artery disease. It was found that the sound generated by saline pulsation through cadaverous hearts was different for diseased and healthy coronary arteries. This difference was indicated by the location of pulsations in the frequency spectrum. There is evidence that these pulsations can be detected by an external microphone on the chest of a patient, and identified by frequency analysis.

References

1. Facts on Major Killing and Crippling Diseases. *Report*, National Health Education Committee, 1959, p. 4.
2. Principles of Internal Medicine (book), Tinsley R. Harrison. "Disorders of the Cardiovascular System," McGraw-Hill Book Company, Inc., New York, N. Y., 1958, pp. 1235–36.
3. Pathology of the Heart (book), S. E. Gould "Coronary Blood Flow," Charles Thomas, Springfield, Ill., 1953, pp. 156–58.
4. Registration and Interpretation of Normal Phasic Inflow into a Left Coronary Artery by an Improved Differential Manometric Method, D. E. Gregg, H. D. Green. *American Journal of Physiology*, Washington, D.C., 1940, p. 121.
5. A General Theory of the Causes of Murmurs in the Cardiovascular System, D. L. Bruns. *American Journal of Medicine*, New York, N. Y., vol. XXVII, Sept. 1959, pp. 360–74.

Technical Solutions to the Problem of Reducing Patient Dosage

J. E. JACOBS
ASSOCIATE MEMBER AIEE

ONE OF THE GREATEST challenges existing today in the field of radiology is the development of equipment and techniques which will permit an accurate diagnosis and at the same time assure that the patient is subjected to the minimum radiation theoretically needed to accomplish that diagnosis. It is the purpose of this paper to discuss the present knowledge concerning what constitutes the minimum dosage and the means available either now or in the not too distant future to produce a diagnostic image utilizing this minimum dosage.

Theoretical Considerations Governing Minimum Patient Dose

There is with X radiation a well-established minimum to the value of patient dosage below which one cannot operate if a diagnostic image is to be retained. This minimum dose still exists even if the perfect image reproducing system is available as a receiver. First recognition of this fact and its subsequent ramifications was given in 1948 by Sturm and Morgan in their now classic treatise on the subject.[1] During the past decade new materials and techniques have placed their ideal

---

Paper 59-236, recommended by the AIEE Electrical Techniques in Medicine and Biology Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Winter General Meeting, New York, N. Y., February 1–6, 1959. Manuscript submitted November 1, 1957; made available for printing October 2, 1959.

J. E. Jacobs is with the General Electric Company, Milwaukee, Wis.

The author wishes to acknowledge the participation of the following people in this project. Particularly helpful in our theoretical studies have been Dr. R. E. Damon, Dr. J. R. Eshbach, Dr. J. L. Lawson, and Dr. P. E. Paschler of the General Electric Research Laboratory staff. In the TVX system and tube fabrication Mr. H. Berger, Mr. C. Hart, and Mr. N. Porath of our laboratory have contributed immeasurably. And the valuable assistance of our consultant, Dr. Rudolf Frerichs, of Northwestern University, Evanston, Ill., is acknowledged.

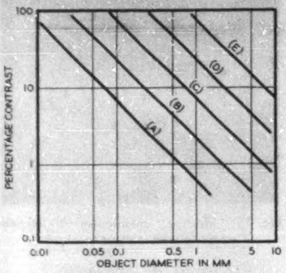


Fig. 1. Theoretical performance of ideal image amplifier if film is used as recording medium

Total exposure to amplifier:
A—10 mr
B—1.0 mr
C—0.1 mr
D—0.01 mr
E—1.0 μr

X-ray image amplifier within the realm of practicality; hence, a brief rèsumè of the considerations underlying this concept will be given.

If an X-ray image reproducing system has infinite brightness gain and is in itself grainless, there still exists a minimum patient dose below which a satisfactory diagnostic image cannot be obtained. This behavior is the result of the statistical fluctuations in the X-ray beam itself. Thus, by examining the effects of these fluctuations on the information content of the resultant image it is possible to predict precisely the minimum obtainable patient dose for a variety of techniques, both in fluoroscopy and radiography. Techniques using patient dosage levels below this theoretical minimum can only be used with a subsequent reduction in available detail in the resultant image. The point is rapidly being reached in the application of X-ray image reproducing systems where such a decision of diagnosing at reduced detail must be made by the radiologist.

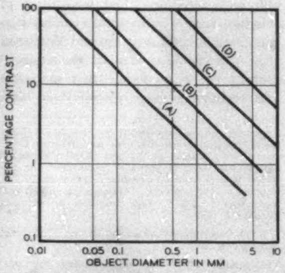


Fig. 2. Theoretical performance of ideal fluoroscopic amplifier

Dosage at amplifier screen:
A—0.6 r/minute
B—0.06 r/minute
C—6 mr/minute
D—0.6 mr/minute

According to statistical theory the ability to detect an object of given size depends not only on the contrast of that object with its background, but also on the total information, or in this case quanta, considered in the time allotted for detection.[2] If one takes film as a recording medium and rules out the necessity for seeing motion, then assuming a signal-to-noise ratio of five a chart of object size as a function of percentage contrast with the surroundings may be drawn. This chart is shown in Fig. 1 and gives the performance of the ideal X-ray image reproducing system, provided that the image is recorded on film. From this, then, may be established the absolute minimum patient dosage to be used if it is desired to see a given detail at a given contrast. The dosage given in Fig. 1 is that at the image reproducing device and for all practical purposes is the patient exit dose. The entrance or skin dose depends on patient thickness and technique used.

The advantage of high-voltage techniques in reducing patient dosage is immediately apparent from these considerations. In fluoroscopy the dosage necessary to create the image must be given during the integration or recognition time of the eye. This assumes that the image amplifier supplies sufficient gain and detail to make the quantum nature of the X-ray beam the limiting factor. A commonly accepted value for this accumulation time of the eye is 0.2 second, althugh studies of radar target recognition have indicated that the eye is capable of integrating weak signals over periods of time approaching tens of seconds. Using the established figure of 0.2 second, a second table of performance for the ideal fluoroscopic amplifier may be constructed and is given as Fig. 2. For cineradiography the dosage of Fig. 1 must be delivered during the frame time. If the image amplifier that is used has limited detail then only the dosage necessary to produce that detail need be given. Any dosage in excess of that is not needed and should not be given. The foregoing considerations assume that the output brightness of the intervening image amplifier is sufficient to expose completely the recording film during the frame time.

Performance of the Ideal Image Amplifier in Radiography

One of the applications of the image amplifiers under development will be as a replacement for the intensifying screens currently used in cassettes. This "electronic cassette," as it has been termed, will permit radiography at a reduced patient dosage. How much this reduction will be depends on how much detail of the resulting image the radiologist is willing to sacrifice. It is difficult to answer this at the present time because the final word depends on many variables, as yet undetermined, although considerable work along these lines is being undertaken. It is interesting to see how the present radiographic screen-film combination compares with the ideal.

An exposure of roughly 1 milliroentgen (mr) is needed to produce a density 1 in the conventional radiographic process. This corresponds roughly to 4×10^4 quanta/mm (millimeter)^2, and hence the curve shown in Fig. 3(A) may be drawn representing the performance of the current radiographic process. Detail is limited by phosphor grain size below 0.1 mm. If the exposure is reduced by ten times, and through the use of electronic image intensification the density of the resultant film is kept at 1.0, then Fig. 3(B) is obtained. It is apparent that even a reduction of dosage to 1/10th that currently used in radiography has serious consequences. Said another way, in the heavy body sections where improvement of the radiographic process is most needed, little hope exists for a reduction in patient dose below 1/2 that currently in use, unless reduced detail is accepted. The amount by which the patient dosage is

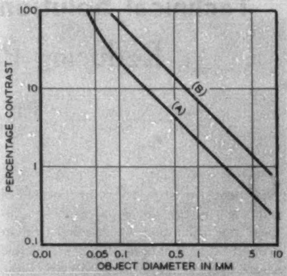


Fig. 3. Comparison of radiography using ideal amplifier intensifying screen having gain of ten times

A—Conventional radiography
B—Radiography using ideal image amplifier of gain 10

reduced thus depends on the ability of the radiologist to work with less detail.

The Ideal Amplifier in Cineradiography

It is in cineradiography that the image amplifier is destined to find widespread use. Two factors are involved. The first is that the brightness gain of the system may be used to its fullest to assure proper density film. The second, and probably most important, is that for motion studies a considerable reduction in detail may be accepted. In fact, each time the necessary detail perception is reduced by a factor of three the patient dosage may be reduced by a factor of ten.

The Ideal Amplifier in Fluoroscopy

The fluoroscopic process is at best inefficient, and expensive from the standpoint of patient dosage. It is will established that a patient dosage reduction of not over 30 times may be achieved with the perfect amplifier before detail suffers. Beyond that point the remainder of the gain of the amplifier goes towards reducing the need for dark adaptation. By use of image amplifiers the fluoroscopic process may be made to approach that of the radiographic one, but only if certain compromises in rate of image presentation are made.

As noted under the previous two sections, a given exposure to the patient is needed to produce an image of a given detail, even if the ideal image producing system is used. If a given number of different images per unit of time is desired, the patient must be exposed to the dosage required to produce those individual images in that length of time. This may be easily verified by a consideration of the dosage at the screen during routine fluoroscopy. This level lies between 10–60 mr/minute.[3] During this time the eye acts as though it made roughly 60 exposures having a detail perception of close to 1 line/mm, providing the object has 100% contrast. The dosage per exposure thus is close to 1 mr. Consider the radiographic process under these conditions where it is found that the dosage per exposure may be as low as 1 microroentgen (μr). Thus an electronic system may be designed to present images of fluoroscopic detail at any repetition rate desired. If one picture per second is chosen then the dosage to the patient may be reduced by a factor exceeding 1,000. If more detail, or a faster presentation rate, is desired, the reduction in dosage to the patient suffers, and the limit will approach the limiting factor of approximately 1/30th that currently used in fluoroscopy.

Commercially Available X-Ray Image Amplifiers

The need for a brighter presentation of the X-ray image has long been recognized and many methods have been proposed, demonstrated in the laboratory, and some made commercially available.[4–6]

Probably the first amplifier proposed, and certainly the first made commercially available, was the electronic image converter tube. This tube, one version of which has been commercially available since 1948, utilizes a fluorescent screen as the primary image converter stage. The light thus produced is used to eject photoelectrically an electron image which in turn is accelerated through an electric field, condensed, and impinged on another fluorescent screen to produce a greatly enhanced image of the X-ray pattern. Brightness gains in excess of 300 times are available in these tubes.

A detailed study of the performance of this approach has been made to show that the fluctuations of the X-ray beam is the limit to obtainable dosage reduction.[7] Amplifiers operating along these lines have become the industry standard.

Another system which utilizes the fluoroscopic screen as the image converter is one in which the image on the screen is detected by a modified studio television camera.[8] This system has the advantage that the brightness of the image presentation is limited only by that of the reproducing Kinescope. Since the electric signal is amplified electronically, numerous schemes of contrast enhancement may be applied to the signal before presentation.

Recently a combination of the image converter tube and television system has been made available. This approach combines the desirable attributes of both systems with the slight disadvantage of additional complexity and bulk.

New Methods of Approaching Theoretical Limit of Minimum Dosages for X-Ray Image Reproduction

The new methods currently in the laboratory or undergoing field evaluation that hold promise of reducing the patient dosage to the theoretical limit may be divided into four classes as follows:

1. Improved fluorescent materials.
2. Increased speed of film emulsions.
3. Solid-state image amplifiers.
4. Electronic systems using photoconductive X-ray camera tubes.

IMPROVED FLUORESCENT MATERIALS

Over the past 50 years considerable effort has been expended in the development of the currently used fluorescent and intensifying screens. It is not surprising therefore, to find that these materials place the radiographic process within 50% of the theoretical limit of minimum patient dosage as set by the fluctuations present in the X-ray beam itself.*

Recently a report was given on intensifying screens using potassium iodide crystals which increased the speed of the process by a factor of four.[9] Reduced detail of the resultant image was reported as one would expect from the considerations previously outlined. Further evaluation will tell whether such a detail reduction is necessarily a detriment to diagnostic films.

INCREASED SPEED FOR X-RAY EMULSIONS

By a control of grain size and orientation, commercially available emulsions of standard film have been increased in speed by a factor of ten. There is an indication this may go to a factor of 1,000. This increase in speed has been made only by sacrificing some detail. It is intriguing to consider the possibility of carrying the techniques into X-ray emulsions. This apparently is not feasible, however, since the X-ray emulsions do not exhibit the same reciprocity as that encountered in the light film. It appears that within the next few years by extensive research the speed of X-ray emulsions may increase a factor of two over that available today. This improvement will place the radiographic process at the theoretical limit of detail.*

SOLID-STATE IMAGE AMPLIFIERS

The solid-state amplifier is unquestionably the hope of the future as far as convenience of use and, at the same time, reduction of dosage to the patient.[10] It may take any one of several forms. These are as follows:

1. Direct conversion of X-ray image to light image.
2. Amplification of light image produced by fluorescent screen.
3. Absorption of X-ray image in photoconductive layer which in turn controls electroluminescent layer which produces light image.

---

* Within the past few years improved emulsions and processing techniques have enabled this factor to be removed, so that currently the radiographic process is at the theoretical limit on available detail.

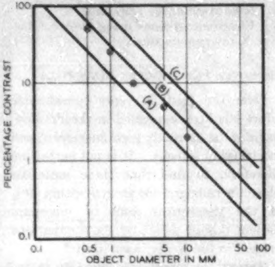


Fig. 4. Measured TVX performance compared to ideal amplifier under conditions as indicated

A—Ideal amplifier frame time 0.2 second; radiation level at tube face 0.01 r/minute.
B—Ideal amplifier frame time 1/30 second; radiation level at tube face 0.01 r/minute.
C—Noise level of amplifier equivalent to 0.002 r/minute.

The last form has received extensive theoretical treatment which indicates that when perfected this form of amplifier will exceed in gain the image converter systems now in widespread use by a factor as high as 100,000.[11] It is obvious from the previous considerations that not all of this gain may be used; however, the added gain of these devices will undoubtedly permit fluoroscopy at dosages 1/30th that in current use, while at the same time eliminating the need for partial dark adaptation. The physical form of these amplifiers is similar to that of the conventional screens and therefore will require a minimum of modification to replace the currently used screens. The only reports given on these amplifiers state that they require a total dose of 25 mr to produce an image.[12] The image once produced will persist until erased. There is, then, at this state of development, a form of an electronic radiographic system the sensitivity of which is essentially that of Blue Brand film used without intensifying screens. The resultant image is produced at a light level of several foot-candles and easily lends itself to viewing in a room having subdued lighting. The brightness gain over a conventional Patterson *B* screen is around 100 times. The inherent limitation on speed of response lies in the photoconductor and will undoubtedly be improved as time goes on.

There is no question at this time that this form of amplifier will eventually become the standard replacement for the fluoroscopic and intensifying screens in the radiographic process. There is no theoretical barrier to its function other than the quantized noise of the X-ray beam. Only the current lack of knowledge of materials processing stands between it and widespread application.

Electronic Systems Using Photoconductive X-Ray Camera Tubes

This system consists of a photoconductive layer that converts the incident X-ray image directly into a charge pattern that is subsequently scanned and presented for viewing on a Kinescope.[13] In 1959 this system, termed TVX, underwent field test in six selected medical clinics. During this time considerable data relating to photoconductor behavior and performance under X radiation were obtained. This information is of great value since it may be extrapolated into anticipating performance of the photoconductor in the solid-state amplifier.

The television-like presentation of X-ray images has many advantages over that of a direct image converter system. Some of the principal ones are:

1. Unlimited brightness of display, hence cineradiography may be performed without the need for increased patient dosage.
2. Image may be viewed remotely at a large number of stations in full daylight.
3. Since the image is available as an electric signal various techniques of electronic contrast enhancement may be easily applied.
4. Through the use of storage techniques a single exposure may be made and the resultant image viewed over a prolonged period of time.
5. A millisecond X-ray generator may be controlled by the unit in such a manner as to produce the exposure during the scanning retrace time. The inherent storage of the pickup tube permits the display of this image at normal rates.
6. The image may be recorded on magnetic tape and replayed at a convenient time.
7. By slow scanning techniques images may be transmitted over local telephone lines.

These are but a few of the advantages; the following are some of the disadvantages of the television presentation of X-ray images.

1. Electronic circuitry is required.
2. Image exposure time is close to 1/30th of a second.
3. Inherent noise level of electronic circuitry limits lowest level of X radiation that may be detected.

The latter two points should be discussed as these are the limiting features of the television presentation when used in general fluoroscopic techniques. The first and most important is that at the brightness levels encountered in a television viewer, i.e., 20–50 foot-candles, the image presentation rate must exceed 48 fields or exposures per second if a flickerless image is to be obtained. For convenience of synchronization and design a 30-frame 60-field interlaced presentation is chosen which conforms with the American Standards of a 525-line picture. This means that during a 1/30th of a second sufficient quanta or dose must be presented to the pickup tube to constitute an image of given detail. Fig. 4 shows the performance of an ideal image amplifier working with the 1/30-second frame time, as well as the curve for the ideal fluoroscopic amplifier using 0.2 second, the integration time of the eye, as the frame time.

As previously mentioned, the noise of the electronic amplifiers limits the lowest level of the usable detected signal in the TVX system. The noise corresponds to a radiation level of 0.002 r (roentgen)/minute as shown on Fig. 4. This means that there exists a threshold of radiation level below which such a system will not operate, which is principally due not to the limitations of the quantized nature of the X-ray beam but more to the noise level of the electronic amplifiers. Armed with this information we are now in a position to evaluate the performance of the systems currently in the field. One is also able to see in what manner this system performance may be improved.

For purpose of comparison, a signal-to-noise of five was chosen which resulted in the radiation level at the tube face being 0.01 r/minute. With the level adjusted in each case to this value the contrast sensitivity of the system was measured as a function of object size. The values obtained are shown on Fig. 4 by the indicated points. It is immediately apparent that the TVX system is performing close to the limit set by X-ray noise as indicated by the proximity of the points to the curve for the ideal fluoroscopic amplifier. It should be recalled that the ideal fluoroscopic amplifier computations were based on a signal to noise ratio of five. For cineradiography, as indicated by the curve for 1/30th-second frame time, the performance seems to exceed this ideal. This obviously cannot exist and is shown not to be the case if the image is recorded on film at 1/30 of a second, and the resultant exposed film examined. Here one sees the buildup and decay of the image in the photoconductor which has in effect increased the effective storage time of the system to a value close to 0.1 second. At higher radiation levels, e.g., 0.1 r/minute, the speed of the system approaches that of the 1/30th-second frame time.

How, then, may the apparent performance of the system be improved as far as a further reduction in patient dosage is concerned? First and most obvious is to decrease the noise level of the electronic amplifiers and/or operate at a lower signal to noise ratio. Both of these factors will help; however, there does not appear at this time to be practically more than a 3-decibel improvement in signal-to-noise ratio on a combination of these schemes.

The most obvious way to reduce dosage and at the same time improve the speed of response of the system is to pulse the X-ray source and utilize the storage feature of the photoconductor in the pickup tube. Preliminary experience indicates that by exposing the pickup target during the scanning beam retrace a reduction in dosage needed at the tube face for a given detail perception of up to three times may be obtained. Since the exposure is in the order of 1-millisecond duration, one has in effect a 1,000 frames/second camera taking radiographs at 1/60-second intervals. For cineradiography this combination approaches the ultimate in performance and simplicity.

Conclusions

Recent technological advances in the field of image intensification systems have reached the point where in the systems available to the radiologist the detail of the resultant image is limited by the quantized nature of the X-ray beam itself. The only avenue open to the radiologist desiring still further reduction in patient dosage is to assure himself that he is not using image detail in excess of that needed for an accurate diagnosis.

With the availability of the solid state image amplifier as a clinical tool this consideration will be paramount in controlling patient dosage.

References

1. Screen Intensification Systems and Their Limitations, R. E. Sturm, R. H. Morgan. *The American Journal of Roentgenology and Radium Therapy*, Springfield, Ill., vol. 62, no. 5, Nov. 1949.

2. Sensitivity Performance of the Human Eye on an Absolute Scale, A Rose. *Journal of the Optical Society of America*, New York, N. Y., vol. 38, 1948, p. 196.

3. Medical Physics (book), Otto Glasser, editor. Year Book Publishers, Inc., Chicago, Ill., 1947 edition, p. 1300.

4. Fluoroscopic Image Brightening by Electronic Means, J. W. Coltman. *Radiology*, Syracuse, N. Y., vol. 51, 1948, p. 359, 337.

5. Amplifying and Intensifying Fluoroscopic Images by Means of a Scanning X-Ray Tube, R. J. Moon. *Science*, Eindhoven, Netherlands, vol. 112, 1950, pp. 389–95.

6. Sichtbarmachung von Roentgenbildern mittels einer auf Roentgenstrahlen ansprechenden Fernsehaufnahmeroehre, M. Richter,¹ M. Pioke. *Zeitschrift für Angewandte Physik*, Berlin, Germany, vol. 7, 1955, p. 562.

7. Electronic Intensification of Fluoroscopic Images, M. C. Teves, T. Tol. *Philips Technical Review*, Eindhoven, Netherlands, vol. 14, no. 2, Aug. 1952, pp. 33–68.

8. The John Hopkins Fluoroscopic Screen Intensifier, R. H. Morgan, R. E. Sturm. *Radiology*, vol. 57, 1951, pp. 556–60.

9. Clinical Applications of KI Intensifying Screens, M. Ter-Pogossian, T. Greitz. American Roentgen Ray Saciety, Oct. 1957.

10. Unique Phosphors That Amplify Light, D. A. Cusano, F. E. Williams. *General Electric Review*, Schenectady, N. Y., Sept. 1956.

11. Solid State Image Amplifiers, G. Diemer, H. A. Klasens, J. G. Van Santen. *Philips Research Reports*, Eindhoven, Netherlands, vol. 10, no. 6, Dec. 1955.

12. An Improved High Gain Panel Light Amplifier, B. Kazan. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 45, no. 10, Oct. 1957.

13. Performance of a Large Area Photoconductive Pickup Tube, J. E. Jacobs, H. Berger. *Electrical Engineering*, New York, N. Y., vol. 75, Feb. 1956, pp. 158–61.

Maximization and Minimization of Complicated Multivariable Functions

NATHAN S. BROMBERG
NONMEMBER AIEE

THIS PAPER is intended as a review and appraisal of various existing methods of finding singular points (maxima, minima, or saddle points) in functions of many variables. It is limited to methods involving a search procedure rather than a formal analytical technique. These methods are useful whenever an optimum condition is required of a system too complex for conventional analytical techniques.

The guiding philosophy is that analytical methods become unwieldy to mechanize in the face of an n-dimensioned problem, and that a procedure based on guesses and subsequent refinements is simpler to mechanize. The approach of converging upon a solution by iteration is also an efficient way to phrase the problem for a digital computer.

These methods are most valuable when the number of variables is large and the functions are nonlinear. Several families of solutions are usually admissible, since most of these problems are not completely constrained. The problem is that of choosing a best solution based on some criterion. These methods may also be applied when a complex system is used as its own model and control laws are desired.

Specifically, we are looking for the maxima or minima of some function in n-dimensional space. We would like to choose successive points, in the surface defined by this function, such that a succession of points leads us to a singular point; hence the general name for these methods, "surface searching." The paper will first consider as an introductory example a linear problem, then the near-quadratic surfaces, and finally higher order surfaces.

Linear programming is a method for minimizing a linear equation of n variables subject to m constraint equations among the n variables ($m<n$). The problem is to find an admissible region in the variable spaces (i.e., constraints are satisfied) and then minimize or maximize the linear equation in this region. The solution of this class of problems will occur on the boundary of the admissible region since the singular points in a linear region occur at the boundaries of a line in n-space. This method has been covered in references 1 and 2.

I. Linear Equations—Kacazmarz's Method

The Kacazmarz method is a way to solve large sets of simultaneous linear equations. A point is arbitrarily chosen in the variable space, and a perpendicular to one of the hyperplanes defined by one lin-

---

Paper 61-803, recommended by the AIEE Basic Sciences Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Summer General Meeting, Ithaca, N. Y., June 18–23, 1961. Manuscript submitted July 29, 1960; made available for printing April 25, 1961.

Nathan S. Bromberg is with the Massachusetts Institute of Technology Electronic Systems Laboratory, Cambridge, Mass.

The author wishes to acknowledge the help gained from discussions with the following members of the Process Control Group in the Electronic Systems Laboratory at Massachusetts Institute of Technology: L. A. Gould, A. E. Beecher, T. J. Boyle, W. Kipiniak, R. Rasche; also, from P. Zieger who reviewed and commented on the paper.

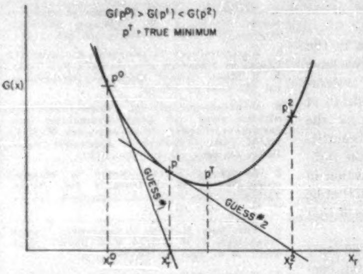


Fig. 1. Overshoot for steepest descent along a tangent

ear equation is drawn. The intersection of this line and the hyperplane (i.e., the foot of the afore-mentioned perpendicular) is used as a starting point for a perpendicular to a second hyperplane. The process is continued in some cyclic order, and the successive points will converge on the point which is the intersection of all the hyperplanes. The nearest possible solution will be approached when the equations are incomplete (more variables than equation). If the number of equations is one less than the number of unknowns then two different initial points will yield different possible solution points. From these two solution points a line may be determined of all possible solutions. The same solution point will, however, be reached if the two points are in a plane perpendicular to the line of all possible solution points. Four non-coplanar points will insure two distinct solution points. This idea is easily extendible to two or more missing equations. See the Appendix for a mathematical discussion and example.

This method, of course, is limited to linear or near-linear functions. The more nearly parallel the equations are, the longer convergence will take. Conversely, perpendicularity between equations will cause immediate convergence.

II. Near-Quadratic Surfaces

A—SOUTHWELL RELAXATION PROCEDURE

Consider a function,

$$G(x_1, x_2, \ldots x_n) = G(x) \qquad (1)$$

The Southwell method consists of computing the first partial derivative of each variable, and selecting the largest partial derivative. A step is then taken in the direction of this variable, and all the first partial derivatives are recalculated at this new point. The rule for step size is a function of the investigator's intuition. The process is repeated and each time the variable whose adjustment will cause the largest favorable change in $G(x)$ is incremented by the magnitude of a step size. For near-quadratic surfaces a good size is given by

$$\epsilon_r = -\left.\frac{\dfrac{\partial G}{\partial x_r}}{\dfrac{\partial^2 G}{\partial x_r^2}}\right|_0 ; \; x_r^{p+1} = \epsilon_r^p + x_r^p \qquad (2)$$

with x_r^p the variable in question, rth dimension, pth iteration.

$G = G(x_1 \; . \; . \; . \; x_n)$ = function to be minimized

$|_0$ means that the derivatives are evaluated at the current point before a new step is taken. See the Appendix for some mathematical discussion of the technique. The better this region approximates a parabolic surface, the better this method is. An external logic must be applied when successive points reverse desirable trends. Such a logic might be a forced direction change or the use of another form for ϵ_r. A second problem occurs when the point about which our parabolic expansion is being taken is an inflection point, since ϵ_r will then be infinite. The breakdown of the Southwell method is due to the failure of this method to treat first-order situations since ϵ_r depends on

$$\frac{\partial^2 G}{\partial x_r^2}$$

very heavily.

B—SOUTHWELL-SYNGE METHOD

This is similar to method II–A. However, the direction r for which

$$\frac{\dfrac{\partial G}{\partial x_r}}{\dfrac{\partial^2 G}{\partial x_r^2}}$$

is maximum is chosen instead of the direction r for which

$$\frac{\partial G(x)}{\partial x_r}$$

is greatest. The relation between this criterion and the latter is developed in the Appendix. This quantity is the expected change in $G(x)$, $\Delta G(x)$, due a step ϵ_r. While the analytic form of step size is the same for the two methods, the actual change in $G(x)$ will only be the same when the surface is exactly quadratic. This method requires more calculation since each

$$\frac{\partial^2 G}{\partial x_r^2}$$

is calculated as opposed to only one

$$\frac{\partial G}{\partial x_r}$$

for method II–B.

C—STEEPEST DESCENT ALONG A TANGENT

In this method a linear approximation is made to the surface

$$G(x) = G(x)\,|_0 + \epsilon_n \nabla G(x)\,|_0 \qquad (3)$$

where $\Delta G(x)$ is the gradient of $G(x)$. Steepest descent along a tangent is very good when $G(x)$ minimum is equal to, or slightly greater than, zero. In the case $G(x) = 0$, at minimum, a step ϵ_r should be taken for each direction r.

From the linear approximation for $G(x)$ and $G(x) = 0$,

$$\epsilon_n = -\left.\frac{G(x)}{\nabla G(x)}\right|_0 \qquad (4)$$

and since

$$\nabla G(x) = \frac{\partial G}{\partial p} = |\text{grad } G(x)| = \left\{\sum_{r=1}^{n}\left(\frac{\partial G}{\partial x}\right)^2\right\}^{1/2} \qquad (5)$$

P is the normal direction to $G(x)$ = constant.

$$\epsilon_r = -\frac{G(x)\cdot\dfrac{\partial G}{\partial x_r}}{\left\{\displaystyle\sum_{r=1}^{n} (G(x))^2\right\}^{1/2}} \qquad (6)$$

and

$$x_r^{p+1} = x_r^p + \epsilon_r^p \qquad (7)$$

The new value of $G(x)$ is calculated using the x_r's, and the process is continued until $G(x) \approx 0$ or slightly larger. If the $G(x)$ minimum is much greater than zero, successive x_r's will start to yield $G(x)$'s whose magnitude oscillates rather than converges. Here the linear fit intersects the line $G(x) = 0$ at a point on the other side of $G(x)$ minimum. (See Fig. 1.)

D—QUADRATIC STEEPEST DESCENT TO A MINIMUM

The steepest descent to a minimum method is more efficient than methods

II-A or II-B. A step is taken in the direction of $\nabla G(x)$ rather then $\nabla G(x_r)$, where the subscript r denotes a particular dimension. Using a parabolic expansion about the present point,

$$G(x) = G(x)\Big|_o + \frac{\epsilon_n}{1!}\cdot\nabla G(x)\Big|_o + \frac{\epsilon_n^2}{2!}\cdot\nabla^2 G(x)\Big|_o \qquad (8)$$

where

$$\nabla^2 G = \frac{\partial}{\partial P}(\nabla G)$$

and P is normal to $G(x) = \text{constant}$. Therefore,

$$\epsilon_n = -\frac{\nabla G}{\nabla^2 G}\Big|_o$$

where the derivation is the same for ϵ_r as in the previous methods. Each ϵ_r may be found from the following, which is easily derivable from vector analysis:

$$\epsilon_r = \epsilon_n \cdot \tau \qquad (9)$$

where τ is the direction cosine

$$\tau = \frac{\dfrac{\partial G}{\partial x_r}}{\left\{\sum_{r=1}^{n}\left(\dfrac{\partial G}{\partial x_r}\right)^2\right\}^{1/2}} \qquad (10)$$

The number of sample points considered here is roughly $1/n$ times the number for methods II-A and II-B. These relaxation methods take the best step in each direction in some cyclic order, while the method of steepest descent goes directly towards the "indicated" singular point. However the amount of calculation necessary for each step in the steepest descent method is greater due to the amount of multiplication and the calculation of all the second-order partial derivatives.

E—Interpolation Scheme[3]

This method tries to overcome the problems of method II-C, namely, the linear extrapolation errors, while retaining the simplicity of calculation in method II-A. After taking an ϵ_n step as in method II-C, a step equal to $1/2\ \epsilon_n$ is taken. Next fit a quadratic to these two points and the present point. The minimum for our fit on $G(x)$ is a function only of the first-order derivatives. The calculation consists of a relatively simple quadratic fit for three points and first-order differentiation. Reference 3 presents some actual tests with these methods. This method breaks down as points are chosen outside of the quadratic fit region.

Concluding Remarks for Near-Quadratic Surfaces

These five methods do not contain operations which discriminate between relative minimums, maximums, and saddle points, and true minimums, and maximums. Of course, the value of $G(x)$ is checked at each step. However, some external strategy is needed to redirect the exploration when the sequence of $G(x)$'s misbehaves (i.e., $G_1(x) > G_2(x) > G_3(x) > G_4(x) < G_5(x) \cdots$). For relative singular points, further exploration starting from new initial conditions is necessary. A saddle point can be checked by a net of trials about the singular point. Once a saddle point has been found, it must be remembered that more than one other singular point is in the vicinity and at least two expeditions from the saddle point should be made. Saddle points will be uncommon finds for these searches. There are, in general, very few space curves which will direct these gradient procedures towards a saddle point.

III. Higher Order Surfaces

Several methods will now be described which are more applicable to exploring high-order hyperplanes. An example is given in reference 4, where the next four methods are compared.

A—Factorial Design

Very simply, this method consists of choosing several values of each factor or variable and then experimenting with all the combinations of the different values for each variable.

Let x_{JK} = the Jth variable and the Kth test value for this variable. The method consists of calculating all the possible

$$G(x_{1K}, x_{2L}, \ldots x_{NM})\text{'s} \qquad (11)$$

where

$K = 1, 2, \ldots k$
$L = 1, 2, \ldots l$
$M = 1, 2, \ldots m$

and there are N variables.

The total number of trials, i.e., times $G(x)$ is calculated, is therefore equal to $(K \cdot L \cdots M)$.

Once this has been completed, either a series of high-order polynomials may be fitted, or the highest observation selected as the minimized or maximized value of $G(x)$. Perhaps a criterion may be established which chooses either the highest observed point or the best value as computed from the fitted curve. The method is easily organized for computation. Inconvenient tests may be postponed. From tests reported in the literature[4] the fitting of curves seems to work except in the region of singular points. It is good for initial exploration of unknown surfaces since the technique is independent of surface geometry. The trial points are predetermined by the experimenter rather than by the results of previous trials.

B—Univariate Method

This is a variation of the relaxation method given in section II-A and II-B. One dimension at a time is explored while the other factors are held constant. A typical experimental design for the univariate method is the following. (1) Choose several equally spaced points along one dimension. (2) Fit a curve to the values of $G(x)$ found by using these points and holding all the other variables constant. (3) Pick the point which maximizes (minimizes) this curve or chose the point in the test set which maximized (minimized) $G(x)$. Note should be made of any other minimums or maximums as evidenced by the curve fit; or other low (high) observations if they are separated from the lowest (highest) result by a hill (valley). This method has the disadvantage that ridges or narrow hills are searched rather inefficiently. It is a computationally simple method. Successive application of the method to all variables seem to confirm the first approximation to the minimum (maximum) of $G(x)$. It is a good method, like method III-A, with which to start a search in order to guess quickly at critical regions.

C—Random Methods

When little is known about a surface or when there are many variables and no reasons to choose specific initial points as is done in the sequential methods (i.e., relaxation gradient), points may be chosen at random. The computation time is a function of the number of points one has time for. To make sure that the surface has been thoroughly searched, a stratified random procedure may be tried. A point is chosen at random within a bounded region. One point may be chosen from each bounded region comprising the surface.

This method is free from being "railroaded" by the surface to relative singular points (i.e., saddle points), but nothing learned in the course of the exploration is used in future trials. The method is useful when a relative singular point has been found and a possibly more fertile region is desired for exploration.

D—Gradient Methods: Descent to Minimum

Let the hypersurface exploration problem be formulated thus[5]:

$$G(x) = G(x_1, x_2, \ldots x_n) \qquad (1)$$

defines a hypersurface, no additional equations (i.e., constraints)

x_j = variables or factors; $j=1, 2, \ldots n$

In general, for analytic solution:

$$\frac{\partial G(x)}{\delta x_j}=0;\ j=1, 2, \ldots n \tag{12}$$

and test for maximum, minimum, or saddle points. Let x^p be a point in the $n+1$ dimension space on the n-dimension hyperplane in question described by the matrix vector:

$$\mathbf{x}^p=\mathbf{u}_1\cdot\mathbf{x}_1+\mathbf{u}_2\cdot\mathbf{x}_2+\ldots\mathbf{u}_n\cdot\mathbf{x}_n \tag{13}$$

where $\mathbf{u}_j$ is a unit vector in jth direction. We now want a sequence of approximations or xP's such that $G_0(x)>G_1(x)>\ldots>G_T(x)$ where $G_T(x)$ is the desired minimum. Now the sequence is specified by the following vector matrix.

$$x]^{p+1}=x]^p+\lambda\cdot d]^p \tag{14}$$

The superscript is the pth iteration. λ is a positive constant and $d]$ is the improvement vector.

All the gradient methods consist of different methods for choosing $d]$ and λ.

The relaxation techniques change one variable at a time before a step is taken, while for gradient methods, a single direction is taken based on the entire group of variables. The relaxation methods usually require more calculation, but they seem to be less complex. The Southwell and Southwell-Synge methods use:

$$d]=\lambda_j\mathbf{u}_j] \tag{15}$$

as j rotates in some cyclic order. This technique requires about n times as many points as the method of steepest descent. The gradient approach uses:

$$d]=[B]^{-1}\cdot\nabla G] \tag{16}$$

where

$$\nabla G(x)]=\left.\begin{matrix}\mathbf{u}_1\cdot\dfrac{\partial G}{\partial x_1}\\ \cdot\,\cdot\\ \mathbf{u}_n\cdot\dfrac{\partial G}{\partial x_n}\end{matrix}\right] \tag{17}$$

$[B]^{-1}$ is an n-by-n metric imposed on the space. Cauchy, who is credited with developing the method of steepest descent in 1847, used

$$[B]^{-1}=[I]=\text{identity matrix} \tag{18}$$

The reason for a metric is to improve the rates of conversion towards a solution. It is, in effect, a changing of the scales in each direction. If $x']=[C]\cdot x]$, then $[B]=[C]_t\cdot[C]$ because $[B]$ must be symmetric and positive definite if distances are to be real numbers.

Newton's method uses:

$$[B]=[L]$$

where the elements of $[L]$ are

$$\frac{\partial^2 G}{\partial x_j\partial x_i};\ j, i=1, 2, \ldots n \tag{19}$$

If $G(x)$ is quadratic, the iteration is completed in one step.

The optimum gradient method is the following procedure: Once a $\nabla G(x)$ direction has been determined, λ is chosen so that $G(x)$ is maximized in this direction. Then $\nabla G(x)$ is calculated again and so on until a singular point is reached. In the method of steepest descent, on the other hand, a step is taken in the $\nabla G(x)$ direction, and at the new point a new $\nabla G(x)$ is calculated, etc. The gradient methods make most of their effort in the regions of singular points.

A λ which is dependent upon $G(x)|_0$ and $\nabla G(x)|_0$ is sometimes used to accelerate or prevent oscillation in a singular region. At random a 1.2 or 0.8 factor may be applied to a constant λ in order to accelerate the finding of a solution. If the convergence misbehaves, halving the step size is a helpful procedure. λ may be a function of the angle between $d]^p$ and $d]^{p+1}$. The larger the angle, the smaller is λ. This scheme requires much calculation to estimate $d]^{p+1}$ at $\mathbf{x}^p$ and $\mathbf{x}^{p+1}$. There are better ways to go on from $\mathbf{x}^{p+1}$ rather than by calculating several new $d']^{p+1}$'s. A rather simple way to accelerate the method is to allow larger round-off errors in calculating the various derivatives and only calculate $G(x)$ with accuracy. This simplifies calculation while adding a "dither" to the procedure. It is necessary to keep a careful watch on the succession of $G(x)$'s in order to be able to apply an appropriate pinpointing method when a singular point is reached.

Noise problems exist since derivative action is employed. If some statements can be made representing the maximum rate at which changes of factors can be expected to occur as the surface is searched, a digital low-pass filter can be used to delete some of the noise. The cut-off frequency would be of the order of one over this maximum rate of change. In other words, can something be said about a maximum expected gradient (a sort of maximum expected rise time)?

In order to test singular points for possible saddle points, examination of the $[L]$ matrix discussed earlier is helpful. $[L]$ must be positive definite for a minimum. For an excellent treatment of the saddle point, see reference 5.

Constraints of the following form will now be considered:

$$f_i(x_1, \ldots x_n)=0;\ i=1, 2, \ldots m \tag{20}$$

and $m<n$ so that the problem is still underconstrained.

It will be helpful to define the following terminology:

$k=n-m$

m dependent variables, $v_1, \ldots v_m$

k independent variables, $w_1, \ldots w_k$

$$x\Big]=\Big.\cdot\cdot\frac{v}{w}\cdot\cdot\cdot\Big]$$

$$\underbrace{\nabla G_w}=\frac{\partial G}{\partial w_1}\cdots\frac{\partial G}{\partial w_k}$$

$$\underbrace{\nabla G_v}=\frac{\partial G}{\partial v_1}\cdots\frac{\partial G}{\partial v_m}$$

$$[v_w]=\frac{\partial v_i}{\partial w_j};\quad \begin{matrix}i=1, 2, \ldots m \text{ rows of matrix}\\ j=1, 2, \ldots k \text{ columns of matrix}\end{matrix}$$

$$\partial v]=\left.\begin{matrix}\partial v_1\\ \cdot\,\cdot\\ \partial v_m\end{matrix}\right];\quad \partial w]=\left.\begin{matrix}\partial w_1\\ \cdot\,\cdot\\ \partial w_k\end{matrix}\right]$$

$$[F_v]=\frac{\partial f_i}{\partial v_j};\quad \begin{matrix}i=1, 2, \ldots m \text{ rows of matrix}\\ i=1, 2, \ldots m \text{ columns of matrix}\end{matrix}$$

$$[F_w]=\frac{\partial f_i}{\partial v_j}\quad \begin{matrix}j=1, 2, \ldots m \text{ rows of matrix}\\ i=1, 2, \ldots k \text{ columns of matrix}\end{matrix}$$

From

$$f_j(v, w)=0 \tag{21}$$

we get

$$df_j]=[F_v]\cdot dv]+[F_w]\cdot dw] \tag{22}$$

In order to obey the constraints we must move only on the lower order hypersurfaces defined by $f_j(v, w)=0$ and $df_j]=0$. Therefore

$$dv]=\ [F_v]^{-1}\cdot[F_w]\cdot dw] \tag{23}$$

$[F_v]^{-1}$ exists if the Jacobian of $[F_v]$ is nonzero everywhere in an open set. If the Jacobian is zero, then the constraints do not implicitly define the dependent variables in terms of the independent ones. If we were able to solve for the dependent variables explicity as in equation 24 we could use the following technique:

$$v_j=f_j(w) \tag{24}$$

Expressing this differentially,

$$dv]=[v_w]\cdot dw] \tag{25}$$

Substituting equation 25 into 23, and canceling $dw]$,

$$[v_w]=-[F_v]^{-1}\cdot[F_w] \tag{26}$$

Therefore

$$\underbrace{\nabla_f G}=\nabla G_w-\nabla G_v\cdot[F_v]^{-1}\cdot[F_w] \tag{27}$$

$$w]^{p+1}=w]^p+\lambda d]^p \tag{28}$$

$$d]^p=-[B]^{-1}\cdot\nabla_f G]^p \tag{29}$$

where $[B]^{-1}$ was the metric imposed on the space, and λ the step size for a method of steepest descent technique.

The constraints, which were satisfied at p, will probably be violated by a small amount at $p+1$. If there is only one constraint equation the following procedure is helpful. From $p+1$ a step is taken perpendicular to the constraint surface and in such a direction that the constraint is satisfied. This procedure will correct the violation by making the shortest digression from $p+1$. Thereafter the surface search is carried out along the boundary until ∇G points back into the allowable region.

Next techniques are considered when there is more than one constraint equation. Let us denote the amount of violation of the constraint by ϵ_j

$$\epsilon_j = f_j(v, w) \tag{30}$$

and expand ϵ_j using the first two terms of a Taylor series about $\epsilon_j=0$.

$$\epsilon_j = 0 + [F_v]\cdot dv] + [F_w]\cdot dw] \tag{31}$$

Let $\epsilon^{p+1}=0$ so that a correcting term $\Delta v]$ may be obtained.

$$\Delta v] = -[F_v]^{-1}\cdot[F_w]\cdot\Delta w] = [v_w]\cdot\Delta w] \tag{32}$$

and

$$v]^{p'+1} = v]^{p+1} + \Delta v]^{p+1} \tag{33}$$

The second technique is based on the idea that a tolerance is allowed for each criterion. For example:

$$|f_j(v, w)| = f_j \tag{34}$$

A strategy can be formulated which tries to minimize the largest f_j as the surface is searched. When a satisfactory point, determined by minimizing (maximizing) $G(v, w)$, is reached the f_j's will be in effect a statement about how well the solution behaves relative to the constraints.

In the case of constraint equations which are given as inequalities, we have exactly the problem above, i.e.,

$$f_j(v, w) \leq Z^2 \tag{35}$$

where Z^2 is the slack variable.

The step size may be a function of the slack variable. Big steps are taken in regions where the constraints are well satisfied, and smaller steps near the constraint boundaries. Fig. 2 is such a strategy. Negative values of λ_m mean that we have crossed over into a nonadmissible region and should turn about.

Lastly, a possible approach is to describe the boundaries of the admissible regions,

$$f_j(v, w) - Z^2 = 0 \tag{36}$$

and proceed to search out the boundary, using $f_j(v, w)$ to constrain this search of the boundary. Having found numerical limits on the variables, the region can now be searched in a manner described earlier in this paper.

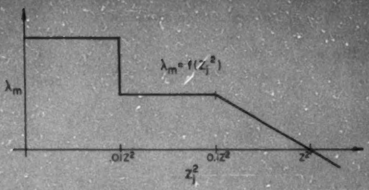


Fig. 2. Step-size strategy as a function of constraint satisfaction

Conclusions

The best scheme would seem to be a combination of the methods presented in this paper. External logic should be able to measure and detect the efficiency of a method in progress and change the search procedure when the efficiency falls below a preset level. Some degree of randomness seems to be required at the start of a search in order to find fertile regions. As the search zeros in on a singular point, an occasional "dither" will prevent oscillations. A series of dithers at a suspected singular point should give some clues as to whether a saddle point, a ridge, or a minimum has been found. The use to which the solutions will be put should help determine the external logic and method. Do we want complete solutions, good starting points? Can past solutions help? These are all important design considerations.

In conclusion it is to be emphasized that iterative methods, as described in this paper, may develop different shortcomings when specific surfaces are searched. Tests should be made on dummy surfaces.

Appendix

Kaczmarz Method[6]—Linear Equations

Variables, x_i; $i=1, 2\ldots n$ co-ordinates orthogonal and same scale in each direction.

(1) u_i = direction number $= y_i - x_i$

(2) $L^2 = \sum_i u_i^2$ $L \geq 0$; L = segment length

(3) Parallel lines $v_i = Ku_i$

(4) Hyperplane; $n-1$ dimensions in n space

(5) Points p on lines through n space determined by following equations:

(a) $\dfrac{x_i - p_i}{u_i} = \dfrac{x_j - p_j}{u_j}$; $j, i = 1, \ldots n$

(b) $\sum_i u_i v_i = 0$ (perpendicular lines, but not necessarily intersecting lines)

(c) Hyperplane representation;

$$\sum_i a_i x_i = b;\ \text{at least one } a \neq 0$$

a and b real numbers

(d) Line segment with $u_i = a_i$ is perpendicular to all lines in hyperplane, whose equation is

$$\sum_i a_i x_i = b$$

and every segment perpendicular to all lines in hyperplane has $u_i \doteq Ka_i$.

A typical use for this method is for large systems of simultaneous linear equations.

Problem

Find a point $(x_1 \ldots x_n)$ at the intersection of the hyperplanes represented by the solution of the equations in the set, i.e.,

$\sum_i a_{ri}\cdot x_i = b_r$; H_r = hyperplane for particular equation

Method

(1) Find a point p_i, such that q_i is at the foot of a perpendicular to H_r from p_i.

(2) Set up the following constraints:

(a) Segment from p_i to q_i perpendicular to all lines in H_r

$$q_i - p_i = Ka_{ri}$$

(b) q_i lies in H_r

$$\sum_i a_{ri} q_i = b_r$$

(3) Iterate in a cyclic order using the la[illegible] q_i^0 for the new p_{i+1}^0 and go to H_{r+1}

$$\sum_i a_{ri}(Ka_{ri} + p_i) = b_r$$

$$K\cdot\sum_i a_{ri}^2 + \sum_i a_{ri} p_i = b_i$$

$$K = \frac{b_r - \sum_i a_{ri}\cdot p_i}{\sum_i a_{ri}^2}$$

Example

(1) $3x_1+2x_2=23$
(2) $-4x_1+x_2=6$

p_β^{α}; α = cycle number; β = hyperplane number

$p_1^0 = 3$
$p_2^0 = 5$
$p_1^1 = 3.92 = q_1^0$
$p_2^1 = 5.61 = q_2^0$

Table I. Successive Approximations

Trial No.	x_1	x_2
0	3	5
1	3.92	5.61
2	0.16	6.54
3	2.33	7.99
4...	0.605...1	8.42...10

$$p_1^2=0.16=q_1^1$$
$$p_2^2=6.54=q_2^1$$
$$p_1^3=2.33=q_1^2$$
$$p_2^3=7.99=q_2^2$$
$$q_1^4=6.05$$
$$q_2^4=8.42$$

Table I gives successive approximations.

The force producing a convergence to a solution is the minimization of a function $f(t)$,

$$f(t)=p_t-q_t$$

so that the following requirements are satisfied:

1. $f(t)$ is mined at desired solution.
2. $f(t)$ is decreased by following the method. The series of steps is convergent.

Southwell Relaxation Technique

In a nearly parabolic region about a solution the following approximation is made to the surface in the x_r direction:

$$G(x_r)=G(x_r)\big|_0+\epsilon_r\cdot\frac{\partial G(x)}{\partial x_r}\bigg|_0+\frac{\epsilon_r^2}{2!}\cdot\frac{\partial^2 G(x)}{\partial x_r^2}\bigg|_0$$

where ϵ_r = step to be taken from present point in x_r direction to minimize $G(x)$. At the minimum we may write

$$\frac{\partial G(x_r)}{\partial \epsilon_r}=0$$

We can solve for ϵ_r with the above two equations.

$$\epsilon_r=\frac{-\dfrac{\partial G(x)}{\partial x_r}}{\dfrac{\partial^2 G(x)}{\partial x_r^2}}\Bigg|_0$$

Southwell-Synge Method

If the change in $G(x)$ due to ϵ_r as given above is calculated, we may arrive at the Synge modification of the Southwell method.

$$\Delta G(x_r)=G(x_r)-G(x_r)\big|_0=\epsilon_r\cdot\frac{\partial G(x)}{\partial x_r}\bigg|_0+\frac{\epsilon_r^2}{2!}\cdot\frac{\partial^2 G}{\partial x_r^2}\bigg|_0$$

$$\Delta G(x_r)=\frac{-\dfrac{\partial G(x)}{\partial x_r}}{\dfrac{\partial^2 G(x)}{\partial x_r^2}}\Bigg|_0$$

and the direction r is chosen for which $\Delta G(x_r)$ is largest.

References

1. Notes on Operations Research (book). The Technology Press, Cambridge, Mass., 1959.

2. Linear Programming: Methods and Applications (book), **S. I. Gass.** McGraw-Hill Book Company, Inc., New York, N. Y., 1958.

3. An Application of the Method of Steepest Descent to the Solution of Systems of Nonlinear Simultaneous Equations, **A. D. Booth.** *Quarterly Journal of Mechanics and Applied Mathematics*, London, England, vol. 2, 1949, pp. 460–68.

4. Comparison of Maximum-Seeking Methods, **S. H. Brooks.** *Operations Research*, Baltimore, Md., vol. 7, July–Aug. 1959.

5. Gradient Methods for the Computer Solution of System Optimizing Problems, **R. R. Brown.** *M.S. Thesis*, Massachusetts Institute of Technology, Cambridge, Mass., 1957.

6. Modern Mathematics for the Engineer (book), **E. F. Beckenbach.** McGraw-Hill Book Company, Inc., New York, N. Y., 1956.

7. The Exploration and Exploitation of Response Surfaces, **G. E. P. Box.** *Biometrics*, New Haven, Conn., vol. 10, 1954, pp. 16–60.

8. A Discussion of Random Methods for Surface Maxima, **S. H. Brooks.** *Operations Research*, Baltimore, Md., vol. 6, 1958, pp. 244–51.

9. Recent Advances in Finding Best Optimum Conditions, **R. L. Anderson.** *Journal*, American Statistical Association, Washington, D. C., vol. 48, 1953, pp. 789–98.

10. Introduction to Numerical Analysis (book), **F. B. Hildebrand.** McGraw-Hill Book Company, Inc., New York, N. Y., 1956, pp. 439–43.

11. Gradient Methods of Maximums, **J. B Crockett, H. Chernoff.** *Pacific Journal of Mathematics*, Berkeley, Calif., vol. 5, 1955, pp. 33–50.

12. Theory of Maximums and Minimums (book), **M. Hancock.** Ginn and Company, Boston, Mass., 1917.

◆

Discussion

Herbert M. Teager (Brookline, Mass.): This paper is an interesting survey of some of the existing gradient methods for locating maxima and minima on surfaces in a dimensional space.

The author considers techniques for surfaces that are linear, quadratic, and of order greater than two. The techniques are somewhat computer-oriented, in the sense that they are iterative rather than direct.

Within the context of the author's basic objective, this is a good paper. There are, however, many important questions that need to be considered in any application of surface-searching techniques to control problems. Among these questions are the following: Is there any way of assuring that the maxima or minima found by these procedures are the best, or near the best? Is there any way of estimating how long the procedure will take to converge, if at all? Assuming that the function is sufficiently well understood so that first and second partial derivatives are known at every point, might it not be possible to employ analytic and heuristic techniques to direct the iterative searches?

Finally, the question of nonanalytic surfaces which do not have the continuity or gradient properties needed for the techniques described by the author might have been considered.

Nathan S. Bromberg: A convenient surface for testing specific procedures is

$$G(X)=\left(\frac{\sin a_1x_1}{b_1x_1}\right)\cdot\left(\frac{\sin a_2x_2}{b_2x_2}\right)\ldots\left(\frac{\sin a_nx_n}{b_nx_n}\right)$$

where n is the number of dimensions. This function $G(X)$ possesses easily located maximum, minimum, relative maxima and minima, and saddle points. The constants a_j and b_k offer sufficient control of $G(X)$ to distort the surface as desired. For example, making $a_1 \gg a_2 \ldots a_n$ will produce a ridge along the x_1 dimension.

I would welcome correspondence relating to practical engineering situations amenable to the surface-searching concept in order to further the work in this area toward more useful techniques.